EGOINTROSPECT: AN EGOCENTRIC DATASET AND BENCHMARK FOR USER-CENTRIC INTERNAL STATE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite extensive efforts on egocentric video datasets and benchmarks, understanding users' internal states, crucial for enabling seamless AI assistant experiences, remains largely overlooked. In this work, we introduce **EgoIntrospect**, the first egocentric dataset captured in user-driven scenarios with self-annotations that explicitly reveal users' interactive intentions with AI assistants. EgoIntrospect was collected using a cross-device setup, providing synchronized video, audio, gaze, motion, and physiological signals. It consists of 180 hours of recordings from 60 subjects, with an average clip duration of 3h. Leveraging EgoIntrospect, we formalize a suite of tasks centered on user internal states, including *affective experience*, *interactive intent*, and *cognitive memory*. We further process the annotations to construct benchmarks that evaluate the ability of modern multimodal large language models (MLLMs) to reason about users' internal states from egocentric observations. Experiments on our benchmark suggest that existing MLLMs struggle to effectively leverage multimodal signals to infer users' subjective internal states. The dataset and annotations will be made publicly available to advance research in egocentric vision and wearable AI assistants.

1 INTRODUCTION

Recent advances in wearable computing (Meta, 2023; Engel et al., 2023) and multimodal large language models (MLLMs) (Yin et al., 2024; Liu et al., 2023; Achiam et al., 2023; Comanici et al., 2025) have enabled a new generation of intelligent, context-aware virtual assistants, capable of operating over multimodal egocentric data and responding to various user-initiated queries. Prior egocentric MLLM works (Lin et al., 2022; Wang et al., 2023; Khirodkar et al., 2023; Ye et al., 2024; Xue & Grauman, 2023; Zhang et al., 2023; Xu et al., 2024) primarily target *external* observation and reasoning—such as analyzing objects, scenes, and user-environment actions. However, real-world AI assistance (Fig. 1) demands understanding users' *internal* states: *how they feel*, *what they need*, and *what matters to them*.

Concretely, we argue that personalized and proactive egocentric AI agents must account for multiple complementary yet interconnected facets of user context: affective awareness captures the user's transient state, interactive intent reflects their immediate needs, and memorability determines which experiences should be retained as personally relevant context. Together, these capabilities constitute a critical yet missing component of user-centric model design, enabling AI agents to adapt to, anticipate, and collaborate with users in everyday settings.

However, the progress towards this goal is bottlenecked by data and benchmarks. While recent work has presented large-scale datasets and benchmarks for objects, actions and scene understanding in egocentric vision (Grauman et al., 2022; 2024), and a small body of work has begun to explore affective experience in this setting (Lin et al., 2024), prior work remains limited in two key aspects. *First*, it typically focuses on a narrow slice of internal state (e.g., emotion), without capturing the broader spectrum of user-centric cognition. *Second*, it predominantly relies on video and audio, overlooking physiological signals that provide complementary and often more direct measurements of a user's internal processes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

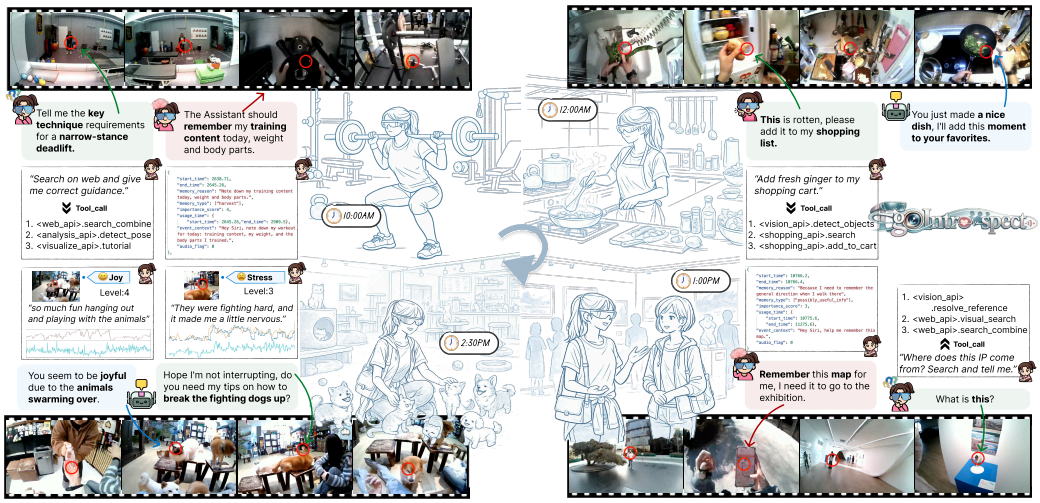


Figure 1: *Visual examples of EgoIntrospect for understanding user internal states*. We illustrate the daily usage of smart-glass AI assistants through three core dimensions: (1) Affective Experience: recognizing salient moments and inferring the user's emotional states (e.g., joy or stress); (2) Interactive Intent: tackling complex, context-grounded queries and initiating proactive assistance (e.g., gym guidance or shopping list management); and (3) Cognitive Memory: modeling memory reachability and the practical need for short-/long-term memory assistance (e.g., navigation guidance or training schedule). Our dataset synchronizes egocentric video, audio, gaze, and physiological signals, bridging first-person observations with the user's internal cognitive processes.

To bridge the gaps, we introduce **EgoIntrospect**, *the first multimodal egocentric dataset and benchmark designed for understanding user internal states in a naturalistic, daily-life setting*. EgoIntrospect is collected using a minimally invasive, wearable platform including everyday smart glasses (Baumann & Dierkes, 2023), an Apple Watch, a physiological watch (Hangzhou Raloway Health Technology Co., Ltd., n.d.), and a smart ring (Tang et al., 2025), capturing synchronized egocentric video, audio, gaze, motion, and physiological and environmental signals. Our dataset comprises over 180 hours of recordings from 60 participants. The key property of EgoIntrospect is that all annotations are provided by users both in-situ and post-hoc, enabling more reliable capture of their affective states, interactive intent, and perceived memorability.

Building on this data, we establish a suite of benchmark tasks that systematically evaluate models' ability to infer and reason about user internal states from multi-modal egocentric data, including:

- *Affective Experience*: probes whether models can identify salient moments users wish to record and infer their emotional status.
- *Interactive Intent*: evaluates whether models can infer user intent by retrieving appropriate external knowledge for a request, and delivering proactive assistance with accurate timing and content.
- *Cognitive Memory*: assesses whether models can estimate what users can vividly remember, as well as what requires memory assistance for future recall.

We note that several tasks are inherently subjective, making reliable evaluation challenging. As detailed in Appendix M, we validate our annotation protocol through a test-retest study and report human-level performance to demonstrate that the resulting task formulations are reliable and solvable. In our experiments, we evaluate 21 state-of-the-art VLMs with modality ablations and an in-context learning analysis. We further examine cross-task interdependence through temporal correlation and transfer experiments, revealing connections among affect, intent, and memory. Due to space constraints, the main paper focuses on the data collection and annotation protocols and benchmark design, with comprehensive experimental results provided in the Appendices G- N.

2 Related Works

Recent years have seen a surge in egocentric datasets. Ego4D and Ego-Exo4D (Grauman et al., 2022; 2024) encompass a vast array of activities. Recently, EgoLife (Yang et al., 2025) collects long egocentric daily life logs, but is limited to a small set of subjects and staged environments. Another

Table 1: *Comparison of egocentric datasets by capture protocol and annotation source.* PS* denotes Physiological Signals. EgoIntrospect differs by adopting user-chosen capture with self-annotation.

Dataset	Domain	Modality	Hours	#Clips	Dur./Clip	Capture Protocol	Annot.
EPIC-KITCHENS (Damen et al., 2022)	Kitchen	Video	100	700	8.5 min	Staged	External
Ego4D (Grauman et al., 2022)	Daily Activities	Video + Gaze + 3D	3,670	9,645	22.8 min	Staged	External
EgoExo4D (Grauman et al., 2024)	Skilled Activities	Video + Gaze + 3D	1,286	5,035	1–42 min	Staged	External
EgoExoLearn (Huang et al., 2024)	Task Execution	Video + Gaze	120	432	13.4 min	Staged	External
EgoLife (Yang et al., 2025)	Daily Life	Video + Gaze + 3D + IMU	266	6	44.3 h	Staged	External
EgoIntrospect	Daily Life	Video + Gaze + **PS***	180	60	**2-4h**	**Subject-chosen**	**Self-Annot.**

Table 2: *Comparison of EgoIntrospectQA with related egocentric VQA benchmarks.* Prior benchmarks primarily target grounding (G), forecasting (F), planning (P), and causality (C), whereas EgoIntrospectQA focuses on user internal states: Affective Experience (A), Interaction Intent (I), and Cognitive Memory (M). *Our benchmark supports multiple formats, including MCQ, binary classification (BC), and multi-class classification (MC).

Benchmark	Source	#QAs	GT Source	Question Type	Task Type
EgoSchema (Mangalam et al., 2023)	Ego4D	5,063	LLM+Annotator	MCQ	Causality
EgoPlan-Bench (Chen et al., 2026)	Ego4D & EPIC-Kitchens	4,939	LLM+Annotator	MCQ	Planning
EgoThink (Cheng et al., 2024a)	Ego4D	700	Annotator	OE	G&F&P
EgoTaskQA (Jia et al., 2022)	LEMMA	40,000	Annotator + Program	OE	G&F&C
EgoMemoria (Ye et al., 2024)	Ego4D	7,026	LLM	MCQ	Grounding
HourVideo (Chandrasegaran et al., 2024)	Ego4D	12,976	LLM+Annotator	MCQ	G&F&C
EgoLifeQA (Yang et al., 2025)	EgoLife	3,000	LLM+Annotator	MCQ	G&C
EgoIntrospectQA	EgoIntrospect	4,060	**Subjects**	**Multiple***	**A&I&M**

related line of work collects physiological signals alongside egocentric data for emotion analysis in lab-controlled environments (Jammot et al., 2025; Park et al., 2020) or for energy expenditure monitoring (Nakamura et al., 2017). Compared to prior efforts, *EgoIntrospect* records multi-sensory long-form streams and adopts a free-form capture protocol defined by participants. And to address the inherent subjectivity, we follow established practices in human-computer interaction (Doherty & Doherty, 2018; Lukka et al., 2024) by adopting both in-situ triggering and retrospective self-annotation, mitigating known biases such as peak-end effects, and salience-driven recall.

Recently, a few egocentric VQA benchmarks have been proposed to evaluate models' ability to reason about first-person behaviors. These recent benchmarks are largely task-driven, where questions are generated from Ego4D or Ego-Exo4D to probe model capabilities in reasoning (Mangalam et al., 2023), grounding (Huang et al., 2025), planning (Peng et al., 2025), and text understanding (Zhou et al., 2025). In contrast, our benchmark follows a user-driven paradigm, focusing on whether MLLMs can interpret users' internal states for seamless AI interactions.

We present additional discussion on related works in Appendix D.

3 Data Collection and Annotation Protocol

3.1 Recording Setup, Participants and Ethics

Devices and data modalities. As shown in Fig. 2, our device setup combines egocentric smart glasses, a wristband, a smart watch, and a smart ring to capture synchronized multimodal signals across different body parts. The Pupil Labs Neon Glasses (Baumann & Dierkes, 2023) record first-person video, eye tracking, audio, and IMU signals of head motion. The wristband records rich physiological and environmental signals. The Apple Watch captures daily activity and mobility statistics, while the smart ring provides fine-grained finger-centric physiological and motion sensing.

Cross-device synchronization. All other data streams were synchronized to the reference time of the smart glasses' video recording. At the start of each recording, participants performed three hand claps within the glasses' field of view. This procedure was repeated at the end of the recording. By visualizing the IMU data into waveforms, annotators identified the corresponding clap timestamps for all devices (Apple Watch, wristband, and ring). With the alignment process, we could retrieve synchronized sensor data through linear interpolation for given video segments.

Participants and Ethics. We recruited N=60 participants through the local campus and social media. Participants received $100 compensation for completing the study. The study was approved by the

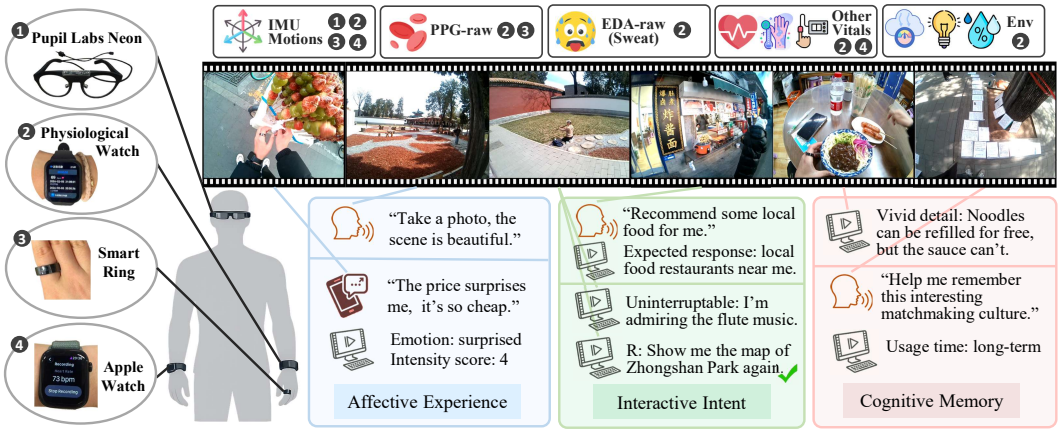


Figure 2: *Capture and annotation overview of EgoIntrospect.* Our dataset records users' natural daily routines with a multimodal wearable sensor suite. We synchronize exteroceptive (video, audio) and interoceptive (physiological, gaze, motion) signals for a rich multimodal capture. Annotations are obtained through two stages: an In-situ Labeling stage, where participants verbally mark key moments or send instant messages during recording, and a Retrospective Annotation stage, where participants subsequently review and refine annotations on our platform.
university's Institutional Review Board (IRB). During consent, participants were informed about the types of data collected by the wearable sensors, the purpose of the study, the planned use of the collected data, and their right to withdraw.

3.2 Experiment Pipeline

Pre-Recording. We introduce the experiment procedure and confirm the activity schedule during recording with the participant. The introduction includes a briefing on the tasks requiring in-situ labeling but not the tasks that only require retrospective annotation.

Data Collection. At the beginning and the end of the recording, participants are required to clap their hands for device synchronization. During recording, participants naturally follow their daily schedule, wearing our devices. Meanwhile, they need to intermittently complete in-situ labeling for our tasks.

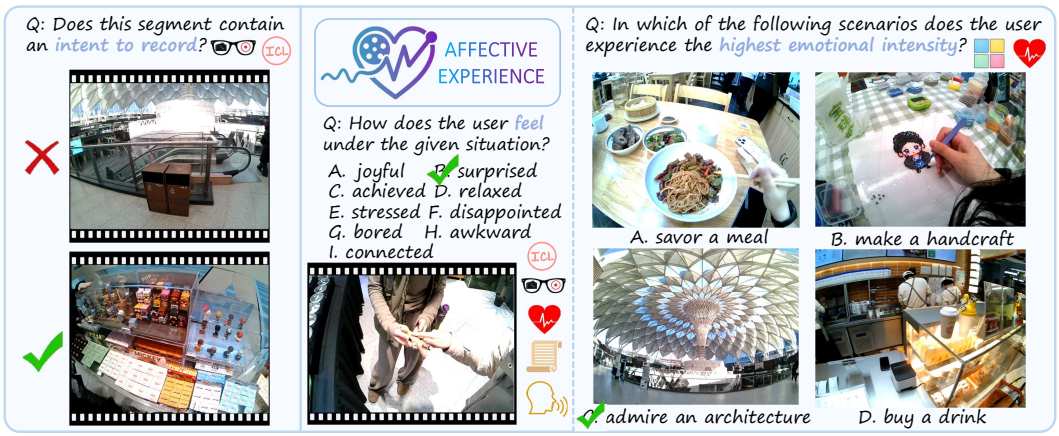
Post-Processing and Annotation. After the collection session, we process the data for annotation. We first export the video recording, compress it, and overlay the gaze trace on the video. We then use ElevenLabs (ElevenLabs) or XunFei (iFLYTEK, 2026) to generate audio transcripts with word-level timestamps, and identify all verbal in-situ commands with GPT-4o (Microsoft, 2026). Within the following week, participants return to the lab and finish retrospective annotations. For detailed procedure, please refer to Appendix C.

3.3 Annotation Protocol

Multi-stage Labeling Process. Our pipeline (see Fig. 2) combines *in-situ labeling* with structured *retrospective annotation* to capture both spontaneous intent and fine-grained semantic information.

- *In-situ Labeling*. Participants provide real-time markers via two approaches to facilitate retrospective annotation: (1) *Voice-based commands*, which are spoken commands starting with *"Hey Assistant"*, used for Task1.1, 2.1, and 3.2; (2) *Instant messages*, where participants send a brief message to the experimenter whenever they noticed a salient emotional response (for Task1.2).
- *Retrospective Annotation*. For tasks with in-situ labels, participants review the extracted message logs, then follow a standardized schema to fill in pre-defined attributes. For tasks without *in-situ* labels, participants review the whole recording to identify and annotate segments for each task. This multi-stage approach ensures that all subjective experiences are saved into structured data.

Data Annotation Platform. Different from previous works, EgoIntrospect's annotations for each user-centered task are provided by the participants themselves. We built a video annotation platform for retrospective annotation by modifying Vidat (Zhang et al., 2020). The platform contains video playback and manipulation functions, as well as task-specific annotation UI. Additional details on the annotation interface and guidelines can be found in the Appendix A.



(a) Moment of Recording: MoR-video (b) Emotion Analysis : EI(left) & EIR(right)

Figure 3: *Illustrative examples of Affective Experience tasks.* (a) MoR-Video/-Photo determine whether a video segment or photo reflects the user's capture intent. (b) Emotion Identification (EI) classifies the user's affective state, and Emotion Intensity Recognition (EIR) selects the scenario with the highest emotional intensity. Icons indicate the specific experimental settings: (ICL) refers to the use of In-Context Learning examples, and the quad-pane icon (⊞) denotes multi-modal candidates.

Test-Retest Reliability. To assess self-annotation reliability, we conducted a test-retest study with 10% of participants after six months. Despite minor recall-related inconsistencies, participants showed high agreement with their initial annotations when reviewing the original recordings, supporting the reliability of our annotation protocol (detailed in Appendix M).

4 Task Definition and Benchmark Design

4.1 Preliminaries

Egocentric Context. Let v denote a task-specific video clip, s denote the synchronized physiological signal, g denote the user's attention signal, and a denote the audio track or transcription, depending on the target model's ability to understand dialogue. Our goal is not to decompose these tasks into task-specific agents, but to evaluate whether a general-purpose MLLM f_θ can infer user-centered states and intentions from egocentric context. Each benchmark question is represented as an instance $\mathcal{I} = \langle \mathcal{X}, q, \mathcal{C} \rangle$, where $\mathcal{X}$ denotes the available multi-modal context, q is the natural-language question, and $\mathcal{C}$ is the candidate answer set. An MLLM receives a serialized multimodal prompt $\Phi(\mathcal{I})$ and predicts an answer: $\hat{y} = f_\theta(\Phi(\mathcal{I}))$. We also experiment with In-Context Learning (ICL) to help the model adapt to individual differences across participants.

Target Space. All tasks are formulated as selecting $y \subseteq \mathcal{C}$ from a candidate set $\mathcal{C} = \{c_i\}_{i=1}^{k} (k \geq 2)$, where c_i represents text, label, object, or multimodal bundle. The ground-truth answer is denoted as $y \subseteq \mathcal{C}$, where y can be single-answer or multi-answer depending on the task. The target space is carefully designed to ensure predictability for subjective tasks, as evidenced by the near-human performance reported in Appendices G–L.

4.2 Affective Experience

4.2.1 Moment of Recording

Motivation. Capturing meaningful moments is one of the most prominent applications of AI glasses. However, existing systems require users to consciously identify interesting events and manually trigger recording, which inevitably interrupts ongoing interactions and implicitly places a cognitive burden on the user. Moreover, what is considered meaningful varies significantly across individuals. Therefore, a future egocentric AI system should learn personalized notions of "interestingness" from historical examples and proactively identify such moments from continuous egocentric streams.

Data Annotation. During data collection, participants marked desired capture moments by in-situ commands *"Hey Assistant, take a photo/record a video"* and could optionally state the reason immediately. During retrospective annotation, they reviewed and annotated each command: for photo commands, they drew a bounding box b_i^* around the intended capture target; for video commands, they marked the start and end time $[t_{s,i}^*, t_{e,i}^*]$ of the intended event segment v_i^*.

Benchmark Design. We design two question types corresponding to the two command categories. A sample of this task is shown in Fig. 3(a).

- *MoR - photo.* We crop a video clip v_i around the command timestamp and craft the candidate answer set to be $\mathcal{C}_i = \{c_i^+, c_{i,1}^-, c_{i,2}^-, c_{i,3}^-\}$, where each option is in the form of a frame overlaid with a bounding box, c_i^+ is the frame containing b_i^* and the distractors c_i^- are manually selected. Let the input be $\mathcal{X}_i = (v_i, a_i, g_i)$, the model outputs a single choice $\hat{y}_i = f_\theta(\Phi((v_i, a_i, g_i), q, \mathcal{C}_i))$.
- *MoR - video.* We use the participant-annotated event segment v_i^+ as the positive segment, and randomly sample unlabeled segments from the same full recording as negative segments v_i^-. Let the input be $\mathcal{X}_i = (v_i, a_i, g_i)$, the model outputs a binary result $\hat{y}_i = f_\theta(\Phi((v_i, a_i, g_i), q)) \in \{0, 1\}$.

Related to Existing Tasks. Video highlight detection (Xiong et al., 2019; Badamdorj et al., 2022; 2021; Liu et al., 2022), which aims to identify salient moments from video sequences, has been extensively studied in computer vision. Compared to this task, our benchmark incorporates additional sensor modalities to capture personalized interesting moments, moving beyond generic saliency toward context-aware, user-specific understanding.

4.2.2 Emotion Analysis

Motivation. Understanding a user's emotional state is essential for delivering personalized and context-aware AI experiences. Our formulation requires models to perform situated reasoning over the environment, user context, and interaction history from first-person audiovisual signals augmented with rich physiological signals in naturalistic settings.

Data Annotation. During retrospective annotation, participants reviewed the full recording and marked segments associated with noticeable emotions. For each segment v_i, they provided an emotion label e_i^*, described the reason $e_i^{r,*}$ for the emotion, and rated its self-perceived intensity $e_i^{I,*}$ on a 1–5 scale. The emotion label can be selected from a predefined set or provided as a custom entry.

Benchmark Design. We design two question types for this task. See a sample in Fig. 3(b).

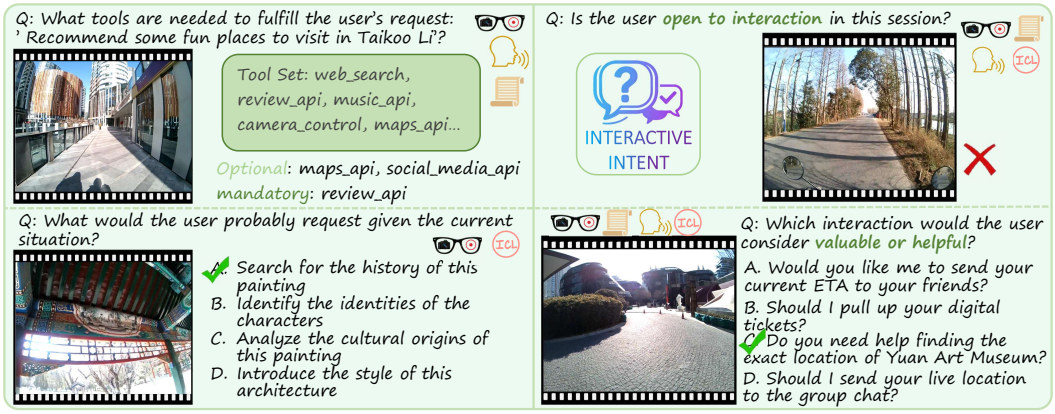
- *Emotion identification (EI).* For each annotated emotion segment v_i, we construct a candidate answer set $\mathcal{C}_i = \{c_i^+, c_{i,1}^-, ..., c_{i,k}^-\}$, where each option c consists of an emotion label and an associated reason. The correct option is $c_i^+ = (e_i^*, e_i^{r,*})$, and the distractor options are generated by pairing other emotion labels with Gemini-generated reasons (The Gemini Team, 2026). Let the input be $\mathcal{X}_i = (v_i, a_i, g_i, s_i)$, the model outputs a single choice $\hat{y}_i = f_\theta(\Phi(\mathcal{X}_i, q, \mathcal{C}_i))$.
- *Emotion Intensity Recognition (EIR).* Instead of predicting the absolute intensity score, this task is formulated as within-user relative comparison. For the same participant and the same emotion label, we sample a set of annotated emotion segments $\{pic_{i_1}, pic_{i_2}, \ldots, pic_{i_k}\}$ (where pic_{i_k} is a key frame) with corresponding intensity scores $\{e_{i_1}^{I,*}, e_{i_2}^{I,*}, \ldots, e_{i_k}^{I,*}\}$. The correct answer is the candidate with the highest or lowest participant-rated intensity. Each candidate in $\mathcal{C}_i$ is defined as $c_{i_j} = (pic_{i_j}, s_{i_j})$. The model outputs a single choice $\hat{y}_i = f_\theta(\Phi(q, \mathcal{C}_i))$.

Related to Existing Tasks. Several recent efforts (Cheng et al., 2024b; Hu et al., 2025; Lian et al., 2025; Lin et al., 2024) have explored emotion analysis with MLLMs. They mainly evaluate whether models can reason about emotions from curated multimodal samples, where emotional states are often inferred from visible expressions or speech. In contrast, our task focuses on the camera wearer's self-reported experience during continuous egocentric recording, where the user's emotion may not be directly observable from facial expressions.

4.3 Interactive Intent

4.3.1 User-Initiated Contextual Request

Motivation. Despite the broad knowledge learned during pretraining, many real-world user requests still require external tools to be fully addressed. To act as an effective assistant, an MLLM must



(a) UCR: TS(above)&RP(below) (b) PRR: PTJ(above)&VII(below)

Figure 4: *Illustrative examples of Interactive Intent tasks*. (a) Tool Selection (TS) identifies the tools needed for addressing the request, and Request Prediction (RP) predicts the user's possible request given the context. (b) Proactive Timing Judgment (PTJ) evaluates the user's openness to interaction, and Valuable Interaction Identification (VII) selects helpful proactive assistance from candidates.

determine what external assistance is needed in a given situation and plan coherent multi-step tool usage to effectively resolve user requests.

Data Annotation. During data collection, participants marked user-initiated requests through in-situ voice commands, using a distinguishable identifier followed by a question or instruction. During retrospective annotation, participants reviewed each request segment v_i and annotated the original request u_i^* and the expected response requirement ρ_i^*.

Benchmark Design. We design two question types for this task, with a sample shown in Fig. 4(a).

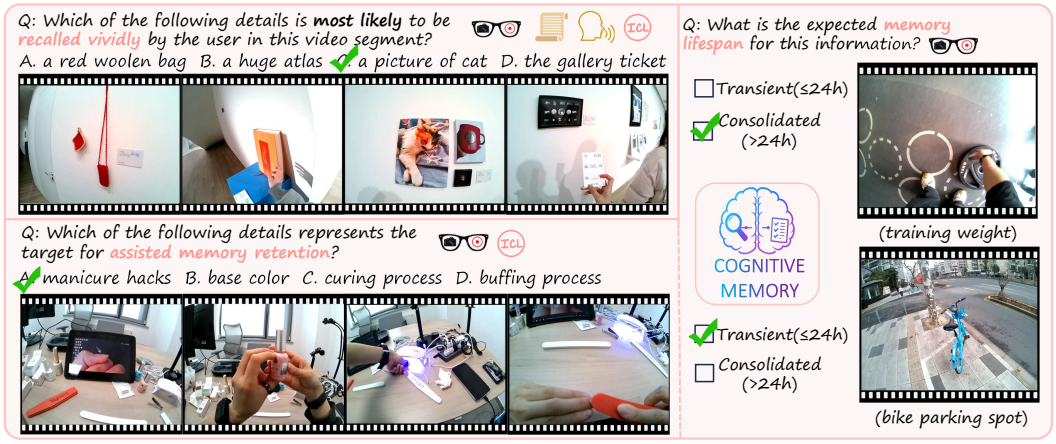
- *Tool Selection (TS)*. We evaluate whether the model can identify what external capability is needed to fulfill the user's request. Given a set of available tools $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_k\}$, we construct the candidate answer set as $\mathcal{C}_i = \mathcal{T} \cup \{\varnothing\}$, where $\varnothing$ denotes that no external tool is required. The ground-truth tool choice is derived by jointly considering u_i^* and ρ_i^*. Let the input be $\mathcal{X}_i = (v_i, a_i, g_i)$, the model outputs one or more choices $\hat{y}_i = f_\theta(\Phi(\mathcal{X}_i, q, \mathcal{C}_i))$.
- *Request Prediction (RP)*. For each request instance, we reformulate request u_i^* and expected response ρ_i^* into a more complete request $\tilde{u}_i^*$, which serves as the correct answer. We construct a four-option candidate set $\mathcal{C}_i = \{\tilde{u}_i^*, u_{i,1}^-, u_{i,2}^-, u_{i,3}^-\}$, where the distractor requests are generated by Gemini. Let the input be $\mathcal{X}_i = (v_i, \tilde{a}_i, g_i)$, where $\tilde{a}_i$ denotes the audio track or transcription with the spoken request removed, the model outputs a single choice $\hat{y}_i = f_\theta(\Phi(\mathcal{X}_i, q, \mathcal{C}_i))$.

Related to Existing Tasks. Existing agent tool-use benchmarks (Qin et al., 2023; Patil et al., 2024) mostly evaluate tool execution or final outcome correctness, whereas our *Tool Selection* question focuses on whether the model can identify the external capabilities needed for a real user request. *Request Prediction* is related to visual question generation (Mostafazadeh et al., 2016), but instead of generating any plausible question from visual content, the model must recover the participant's actual request from captured signals.

4.3.2 Proactive Request Recommendations

Motivation. Unlike reactive assistants that require explicit user requests, proactive assistance enables the system to anticipate user needs based on ongoing context and past behavior, providing seamless and autonomous support. This is particularly important in egocentric streaming settings, where continuous observation allows the model to infer intent and intervene in a timely manner.

Data Annotation. Participants marked forbidden segments where they would not want to receive proactive recommendations, leaving the remaining parts $\{v_i^{int}\}$ as interruptible regions. For these segments, we generated recommendations at a five-minute granularity (detailed in Appendix B.2). Participants reviewed the generated recommendations and selected those they would be willing to accept. For each accepted recommendation $r_{i,j}^+$, they further annotated a presentation segment $v_{i,j}^{rec}$, indicating the time period in which they considered it appropriate for this recommendation to appear.



(a) Memory Recall Prediction(MRP, top left) (b) Memory Intent Modeling : MAR(bottom left)&MLI(right)

Figure 5: *Illustrative examples of Cognitive Memory tasks*. (a) Memory Recall Prediction evaluates the model's ability to identify specific, vivid details that are most reachable in a user's memory. (b) Memory Assistance Recognition (MAR) identifies specific items the user explicitly intended to preserve (MAR-Event is shown; see MAR-Object in text), and Memory Lifespan Identification predicts the temporal utility and retention requirements of a designated memory item.

Benchmark Design. We design two question types for this task. Samples are presented in Fig. 4(b).

- *Proactive Timing Judgment (PTJ)*. We sample positive segments from the interruptible parts and negative segments from the forbidden parts of the same recording. For each sampled segment v_i, let the input be $\mathcal{X}_i = (v_i, a_i, g_i)$ and the binary candidate answer set be $\mathcal{C} = \{True, False\}$, the model outputs a binary choice $\hat{y}_i = f_\theta(\Phi(\mathcal{X}_i, q, \mathcal{C}))$.
- *Valuable Interaction Identification (VII)*. For each accepted recommendation $r^+_{i,j}$, we construct a four-option candidate set $\mathcal{C}_{i,j} = \{r^+_{i,j}, r^-_{i,j,1}, r^-_{i,j,2}, r^-_{i,j,3}\}$, where the distractors are randomly sampled from recommendations rejected by the participants. Let the input be $\mathcal{X}_{i,j} = (v^{rec}_{i,j}, a_{i,j}, g_{i,j})$, the model outputs a single choice $\hat{y}_{i,j} = f_\theta(\Phi(\mathcal{X}_{i,j}, q, \mathcal{C}_{i,j}))$.

Related to Existing Tasks. Several prior works study proactive recommendation with LLMs (Zhao et al., 2024; Gao et al., 2023), but primarily focus on modeling user-preferred topics rather than interaction intent. Proactive interaction understanding with MLLMs remains largely underexplored. Existing work on proactive timing typically aims to identify key visual evidence for VQA (Wang et al., 2025; Zhang et al., 2025b; Qian et al., 2025; Zhang et al., 2025a), whereas our task focuses on modeling internal user states to determine whether an intervention would be disruptive. Moreover, these prior efforts emphasize responding to given queries at the appropriate time, rather than proactively recommending potential interaction options as required in our benchmark.

4.4 Cognitive Memory

4.4.1 Memory Recall Prediction

Motivation. Human memory is inherently selective and prone to decay, creating a functional gap between exhaustive digital logs and subjective recall. While existing systems focus on objective visual search, they fail to account for why specific vivid details persist while entire events become obscured. An AI model that understands human memory mechanisms could construct a hierarchical personal timeline that not only supports cognitively aligned retrieval but also enhances memory recall.

Data Annotation. Post-recording, participants provide retrospective accounts of vivid details they can recall from the session. Based on these, experimenters annotate timestamps for positive instances, while negative distractors are generated by MLLMs from the same video segments.

Benchmark design. This task assesses individual memory reachability by requiring the model to identify a specific, vivid detail that is most likely to be recalled by the user of a past event from a set of plausible distractors. Given the multimodal input $\mathcal{X}_i = (v_i, a_i, g_i)$, a query q, and a candidate set of details $\mathcal{C}_i$, the model must select the correct ground-truth option. The prediction is formulated as: $\hat{y}_i = f_\theta(\Phi(\mathcal{X}_i, q, \mathcal{C}_i)) \in \mathcal{C}_i$. Examples of this task are illustrated in Fig. 5(a).

Related to Existing Tasks. Early work in egocentric vision has explored the discovery of important objects and people for egocentric video summarization (Lee et al., 2012; Lu & Grauman, 2013), yet they do not directly address memorability. While estimating the memorability of images (Isola et al., 2011) and videos (Dumont et al., 2023) has been previously studied, prior work focuses on third-person visual content rather than first-person video. In addition, how MLLMs interpret human memory mechanisms has not been explored in existing benchmarks.

4.4.2 Memory Intent Modeling

Motivation. While the aforementioned memory recall task addresses what users can vividly remember, an AI model must also consider what users want it to memorize. Notably, in continuous egocentric streaming, captured data comprises both passively observed perceptual inputs and salient content that warrants intentional preservation. Developing models that understand user memory needs is therefore critical for reducing cognitive load and transforming passive recording into accurate and effective memory assistance in complex real-world scenarios.

Data Annotation. During recording, participants identify specific recall targets—including events and objects—and assign binary lifespan labels during retrospective sessions. For MAR, v_i contains user-annotated memory intent, with negative distractors generated by MLLMs to ensure competitive difficulty. For MLI, v_i consists of the exact information segment to be memorized, with ground truth labels derived directly from participant annotations.

Benchmark Design. We characterize memory intent through two dimensions: the identification of items to be retained and the temporal relevance. Samples of these tasks are illustrated in Fig. 5(b).

(1) **Memory Assistance Recognition (MAR)**: This task identifies items that a user explicitly intends to preserve for future use. Based on the memory item type, we define the task with two input formats:

- *MAR - Object*. For memory items with bounding box annotation, the candidate answer set $\mathcal{C}_i = \{c_i^+, c_{i,1}^-, c_{i,2}^-, c_{i,3}^-\}$ consists of frames overlaid with bounding boxes, where c_i^+ contains b_i^* and the distractors c_i^- are manually selected. Given the input $\mathcal{X}_i = (v_i, g_i)$, the model predicts a single choice: $\hat{y}_i = f_\theta(\Phi((v_i, g_i), q, \mathcal{C}_i))$.
- *MAR - Event*. For memory items annotated as events, we crop a video clip v_i around the specific timestamps. We then craft textual distractors based on other events occurring within the same segment. With input $\mathcal{X}_i = (v_i, g_i)$, the model outputs a single choice: $\hat{y}_i = f_\theta(\Phi((v_i, g_i), q, \mathcal{C}_i))$.

(2) **Memory Lifespan Identification (MLI)**: This task assesses the model's ability to predict the temporal utility of specific information the user intends to remember. Inspired by the cognitive process of memory consolidation, the model must discern whether a designated memory item represents a transient need or warrants long-term retention. Given the multimodal input $\mathcal{X}_i = (v_i, g_i)$ and a query q regarding a pre-identified memory item, the model predicts a binary classification $\hat{y}_i = f_\theta(\Phi(\mathcal{X}_i, q)) \in \{\mathcal{S}, \mathcal{L}\}$, where:

- $\mathcal{S}$ (**within_24h**): the memory item is relevant only within a short-term, 24-hour window;
- $\mathcal{L}$ (**more_than_24h**): the memory item requires consolidation for retrieval beyond a single day.

Related to Existing Tasks. Episodic memory (Bärmann & Waibel, 2022; Grauman et al., 2022; Datta et al., 2022; Wang et al., 2023; Manigrasso et al., 2026) has been widely studied in egocentric vision. However, prior work typically focuses on evaluating grounding ability by asking models when and where events occur using annotator-defined queries. In contrast, our benchmark adopts a user-centric formulation centered on what users would want to remember for future retrieval.

5 Conclusion

In this paper, we introduce EgoIntrospect, the first dataset and benchmark aimed at advancing personalized egocentric AI assistants. Compared to prior datasets, EgoIntrospect adopts unconstrained capture settings and self-annotation to more reliably capture internal user states. Therefore, our proposed benchmark suite provides a snapshot of how users intend to use wearable AI assistants, highlighting the gaps in current MLLMs. Our work takes an important step toward bridging prevailing MLLMs with AI assistants for wearable devices, and points to exciting future research directions for personalized and proactive AI user experiences.

AI USE STATEMENT

Generative AI systems were used as research tools in several stages of this work. As described in the paper and appendix, speech-recognition services and GPT-family models were used to transcribe recordings and identify in-situ commands; Gemini-family models were used to generate or refine selected benchmark distractors and candidate proactive recommendations; and multimodal language models were evaluated as benchmark baselines or used in explicitly reported auxiliary analyses. All model-generated benchmark content was subject to task-specific constraints and, where it represented a user's intent or preference, was reviewed, selected, edited, or verified by participants or the research team. Generative AI tools were also used to assist with language editing and polish portions of the manuscript. The authors reviewed and revised all AI-assisted text and outputs and take full responsibility for the study design, data, analyses, claims, citations, and final manuscript.

ETHICS STATEMENT

This study involved long-form egocentric video and audio together with gaze, motion, physiological, environmental, and self-reported data. The study protocol was reviewed and approved by the authors' institutional review board. All participants provided informed consent after being told what data would be collected, the purpose and planned use of the data, the compensation, and their right to withdraw. Participants were compensated for their time, and the collection protocol allowed them to exclude content they did not wish to contribute. For privacy-sensitive scenarios involving identifiable social partners, consent for public data sharing was obtained from the relevant individuals; copyrighted-content recording was prohibited.

Because first-person recordings may incidentally contain bystanders, private spaces, screens, speech, addresses, payment information, or other identifiers, public benchmark clips will undergo automatic face blurring followed by manual review. Direct identifiers will be removed or generalized in filenames, metadata, transcripts, and annotations, and clips that cannot be adequately de-identified will not be publicly released. The continuous recordings and aligned multimodal signals will be distributed only through a controlled-access procedure subject to approval and a data-use agreement.

Inferring affect, intent, memory, attention, or physiological state is inherently sensitive and may be inaccurate across individuals and demographic groups. The dataset is intended for research on user-aligned assistive systems, not for clinical diagnosis, surveillance, covert profiling, or consequential decisions about individuals. Any deployed system based on this work should obtain meaningful user consent, minimize data collection and retention, provide user control over sensing and intervention, protect against unauthorized access, and evaluate subgroup performance and failure modes. Our participant pool and geographic coverage remain limited, so the dataset should not be treated as representative of all users or cultures.

REPRODUCIBILITY STATEMENT

The main paper and appendices document the sensing hardware, synchronization and annotation procedures, task definitions, participant-independent train/validation/test split, benchmark construction, prompt templates, evaluation metrics, model input formats, and principal inference settings. To support direct reproduction, we will release the full dataset along with the processed benchmark instances used in the reported experiments, including de-identified task clips, benchmark JSON files, candidate answers, ground-truth labels, prompts, preprocessing and evaluation code, and executable task-specific pipelines.

The complete continuous recordings, synchronized sensor streams, and original retrospective annotations contain substantially more sensitive information and therefore cannot be released without restriction. They will be made available around the conference release through a controlled-access process and data-use agreement after de-identification review. Closed-source API models and preview model versions may change or become unavailable; accordingly, we report the model identifiers, prompts, decoding parameters, frame-sampling or native-video settings, and other available implementation details, while noting that exact reproduction of proprietary-model outputs may depend on provider-side versions and infrastructure.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mohamed Ait Gacem, Saifeddin Alghlayini, Wessam Shehieb, Muaid Saeed, Ahmed Ghazal, and Mustahsan Mir. Smart assistive glasses for alzheimer's patients. In *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 1–5. IEEE, 2019.

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8127–8137, October 2021.

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Contrastive learning for unsupervised video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14042–14052, June 2022.

Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1560–1568, 2022.

Chris Baumann and Kai Dierkes. Neon accuracy test report. 2023. doi: 10.5281/ZENODO.10420388. URL https://zenodo.org/doi/10.5281/zenodo.10420388.

Runze Cai, Nuwan Janaka, Hyeongcheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. Aiget: Transforming everyday moments into hidden knowledge discovery with ai assistance on smart glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–26, 2025.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Yuhu Chang, Yingying Zhao, Mingzhi Dong, Yujiang Wang, Yutian Lu, Qin Lv, Robert P Dick, Tun Lu, Ning Gu, and Li Shang. Memx: An attention-aware smart eyewear system for personalized moment auto-capture. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *International Journal of Computer Vision*, 134(3):118, 2026.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14291–14302, 2024a.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37: 110805–110853, 2024b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19119–19128, June 2022.

Kevin Doherty and Gavin Doherty. The construal of experience in hci: Understanding self-reports. *International Journal of Human-Computer Studies*, 110:63–74, 2018.

Théo Dumont, Juan Segundo Hevia, and Camilo L Fosco. Modular memorability: Tiered representations for video memorability prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10751–10760, 2023.

ElevenLabs. Speech to text. `https://elevenlabs.io/docs/eleven-creative/playground/speech-to-text`. ElevenLabs Documentation. Accessed: 2026-05-06.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*, 2023.

Olga Gelonch, Mireia Ribera, Núria Codern-Bové, Sílvia Ramos, Maria Quintana, Gloria Chico, Noemí Cerulla, Paula Lafarga, Petia Radeva, and Maite Garolera. Acceptability of a lifelogging wearable camera in older adults with mild cognitive impairment: a mixed-method study. *BMC geriatrics*, 19(1):110, 2019.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

Raphaël A El Haddad, Zeyu Wang, Yeonsu Shin, Ranyi Liu, Yuntao Wang, and Chun Yu. Ar secretary agent: Real-time memory augmentation via llm-powered augmented reality glasses. *arXiv preprint arXiv:2505.11888*, 2025.

Hangzhou Raloway Health Technology Co., Ltd. Ralovi S2 Wristwatch: EDA Skin Conductance and HRV Heart Rate Variability. `https://raloway.com/index.php?m=home&c=Lists&a=index&tid=23`, n.d. Accessed: 2026-05-07.

Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. Sensecam: A retrospective memory aid. In *International conference on ubiquitous computing*, pp. 177–193. Springer, 2006.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, and Yu Qiao. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22072–22086, June 2024.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, Ning Zhang, Yong Jae Lee, and Miao Liu. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24169–24179, June 2025.

iFLYTEK. Large Model for Audio File Transcription. `https://www.xfyun.cn/doc/spark/asr_llm/Ifasr_llm.html`, 2026. iFLYTEK Open Platform Documentation Center. Accessed: 2026-05-06.

P Isola, Jianxiong Xiao, A Torralba, and A Oliva. What makes an image memorable? In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 145–152, 2011.

Matthias Jammot, Björn Braun, Paul Streli, Rafael Wampfler, and Christian Holz. egoemotion: Egocentric vision and physiological signals for emotion and personality recognition in real-world tasks. *arXiv preprint arXiv:2510.22129*, 2025.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos, 2022. URL `https://arxiv.org/abs/2210.03929`.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL `https://github.com/ultralytics/ultralytics`.

Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19807–19819, 2023.

Jangho Kwon, Jihyeon Ha, Da-Hye Kim, Jun Won Choi, and Laehyun Kim. Emotion recognition using a glasses-type wearable device via multi-channel facial responses. *IEEE Access*, 9:146392–146403, 2021.

Huy Viet Le, Sarah Clinch, Corina Sas, Tilman Dingler, Niels Henze, and Nigel Davies. Impact of video summary viewing on episodic memory recall: Design guidelines for video summarizations. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 4793–4805, 2016.

Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, et al. Sensible agent: A framework for unobtrusive interaction with proactive ar agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2025.

Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. Gazepointar: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2024.

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1346–1353. IEEE, 2012.

Franklin Mingzhe Li, Di Laura Chen, Mingming Fan, and Khai N Truong. Fmt: A wearable camera-based object tracking memory aid for older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–25, 2019.

Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, François Guimbretière, and Cheng Zhang. Eyeecho: Continuous and low-power facial expression tracking on glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.

Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6731–6747, 2021.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models, 2025. URL https://arxiv.org/abs/2501.16566.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

Wang Lin, Yueying Feng, WenKang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. E^3: Exploring Embodied Emotion Through A Large-Scale Egocentric Video Dataset. *Advances in Neural Information Processing Systems*, 37:118182–118197, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3042–3051, June 2022.

Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2714–2721, 2013.

Lauri Lukka, Veli-Matti Karhulahti, Vilma-Reetta Bergman, and J Matias Palva. Measuring digital intervention user experience with a novel ecological momentary assessment (ema) method, corto. *Internet Interventions*, 35:100706, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Zaira Manigrasso, Matteo Dunnhofer, Antonino Furnari, Moritz Nottebaum, Antonio Finocchiaro, Davide Marana, Rosario Forte, Giovanni Maria Farinella, and Christian Micheloni. Online episodic memory visual query localization with egocentric streaming object memory. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3951–3960, March 2026.

Katsutoshi Masai, Kai Kunze, Yuta Sugiura, Masa Ogata, Masahiko Inami, and Maki Sugimoto. Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(4): 1–23, 2017.

Meta. Introducing the new ray-ban meta smart glasses. https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/, September 2023. Accessed: 2026-04-10.

Microsoft. OpenAI GPT-4o. https://ai.azure.com/catalog/models/gpt-4o, 2026. Microsoft Foundry Models catalog. Version 2024-11-20; last updated April 2026. Accessed: 2026-05-06.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1802–1813, 2016.

Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1868–1877, 2017.

Evangelos Niforatos, Caterina Cinel, Cathleen Cortis Mack, Marc Langheinrich, and Geoff Ward. Can less be more? contrasting limited, unlimited, and automatic picture capture for augmenting memory recall. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(2): 1–22, 2017.

Ye Pan, Chi Kit Wong, Yuanhuiyi Lyu, Hanqian Li, Jiahao Huo, Jiacheng Chen, Lutao Jiang, Xu Zheng, and Xuming Hu. Egointent: An egocentric step-level benchmark for understanding what, why, and next. *arXiv preprint arXiv:2603.12147*, 2026.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, 2020.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 126544–126565. Curran Associates, Inc., 2024. doi: 10.52202/079017-4020. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e4c61f578ff07830f5c37378dd3ecb0d-Paper-Conference.pdf.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *arXiv preprint arXiv:2509.07447*, 2025.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24045–24055, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023. URL https://arxiv.org/abs/2307.16789.

Zhiwen Qiu, Mojtaba Ashour, Xiaohe Zhou, and Saleh Kalantari. Navmarkar: A landmark-based augmented reality (ar) wayfinding system for enhancing older adults' spatial learning. *Advanced Engineering Informatics*, 62:102635, 2024.

Corina Sas, Tomasz Fratczak, Matthew Rees, Hans Gellersen, Vaiva Kalnikaite, Alina Coman, and Kristina Höök. Affectcam: arousal-augmented sensecam for richer recall of episodic memories. In *CHI'13 extended abstracts on human factors in computing systems*, pp. 1041–1046. 2013.

Abigail J Sellen and Steve Whittaker. Beyond total capture: a constructive critique of lifelogging. *Communications of the ACM*, 53(5):70–77, 2010.

Abigail J Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. Do life-logging technologies support memory for the past? an experimental study using sensecam. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 81–90, 2007.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023. doi: 10.1109/ICCV51070.2023.01101.

Anna Ståhl, Kristina Höök, Martin Svensson, Alex S Taylor, and Marco Combetto. Experiencing the affective diary. *Personal and Ubiquitous Computing*, 13(5):365–378, 2009.

Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. Visual intention grounding for egocentric assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2512–2522, 2025.

Jiankai Tang, Kegang Wang, Yingke Ding, Jiatong Ji, Zeyu Wang, Xiyuxing Zhang, Ping Chen, Yuanchun Shi, and Yuntao Wang. A dataset and toolkit for multiparameter cardiovascular physiology sensing on rings. *arXiv preprint arXiv:2505.04172*, 2025.

The Gemini Team. Gemini 3.1 Pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, February 2026. Accessed: 2026-05-06.

Vijay Veerabadran, Fanyi Xiao, Nitin Kamra, Pedro Matias, Joy Chen, Caley Drooff, Brett D Roads, Riley Williams, Ethan Henderson, Xuanyi Zhao, et al. Benchmarking egocentric multimodal goal inference for assistive wearable agents. *arXiv preprint arXiv:2510.22443*, 2025.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=DTvviEnW2A.

Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. *arXiv preprint arXiv:2312.05269*, 2023.

Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. G-voila: gaze-facilitated information querying in daily scenarios. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–33, 2024.

Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. Emo: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, pp. 448–461, 2020.

Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13525–13536, June 2024.

Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in neural information processing systems*, 37:1890–1918, 2024.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28885–28900, 2025.

Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 982–990, 2016.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mm-ego: Towards building egocentric multimodal llms for video qa. *arXiv preprint arXiv:2410.07177*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL http://dx.doi.org/10.1093/nsr/nwae403.

Yuqian Yuan, Ronghao Dang, Long Li, Wentong Li, Dian Jiao, Xin Li, Deli Zhao, Fan Wang, Wenqiao Zhang, Jun Xiao, et al. Eoc-bench: Can mllms identify, recall, and forecast objects in an egocentric world? *arXiv preprint arXiv:2506.05287*, 2025.

Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13901–13912, October 2023.

Jiahao Zhang, Stephen Gould, and Itzik Ben-Shabat. Vidat—ANU CVML video annotation tool. https://github.com/anucvml/vidat, 2020.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12055–12079, 2025a.

Yulin Zhang, Cheng Shi, Yang Wang, and Sibei Yang. Eyes wide open: Ego proactive video-llm for streaming video. *arXiv preprint arXiv:2510.14560*, 2025b.

Yingying Zhao, Yuhu Chang, Yutian Lu, Yujiang Wang, Mingzhi Dong, Qin Lv, Robert P Dick, Fan Yang, Tun Lu, Ning Gu, et al. Do smart glasses dream of sentimental visions? deep emotionship analysis for eyewear devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–29, 2022.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907, 2024.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3363–3373, 2025.

Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024.

APPENDIX OVERVIEW

- **Appendix A**: Annotation Platform(Page 19)
- **Appendix B**: Data Processing Algorithms for Annotation(Page 24)
- **Appendix C**: Detailed Experiment Pipeline(Page 28)
- **Appendix D**: More on Related Work(Page 32)
- **Appendix E**:Dataset and Benchmark Statistic(Page 33)
- **Appendix F**: Dataset Open-Source(Page 37)
- **Appendix G**: Benchmark Task1.1 - Moment of Recording(Page 38)
- **Appendix H**: Benchmark Task1.2 - Emotion Analysis(Page 43)
- **Appendix I**: Benchmark Task2.1 - User-Initiated Contextual Request(Page 45)
- **Appendix J**: Benchmark Task2.2 - Proactive Request Recommendations(Page 49)
- **Appendix K**: Benchmark Task3.1 - Memory Recall Prediction(Page 57)
- **Appendix L**: Benchmark Task 3.2 - Memory Intent Modeling(Page 60)
- **Appendix M**: Annotation Reliability Analysis and Test-Retest Results(Page 67)
- **Appendix N**: Cross-Task Interdependence Study(Page 68)
- **Appendix O**: Limitation(Page 70)

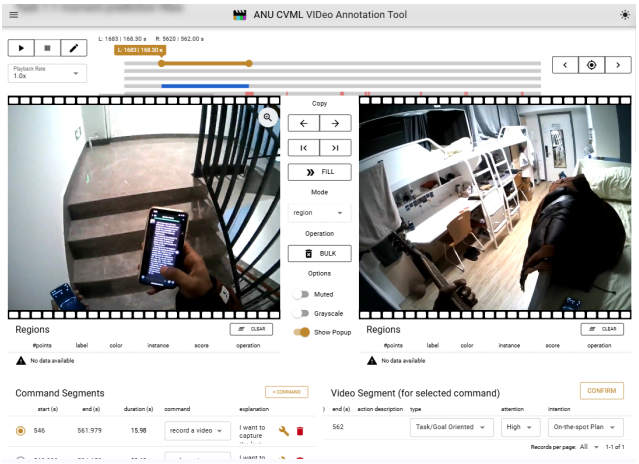


Figure 6: Annotation interface for Task1.1 (Moment of Recording).

A Annotation Platform

A.1 Task1.1 - Moment of Recording

Each participant annotated their egocentric video recordings using a custom web-based tool, Vidat, following a structured retrospective labeling protocol. The annotation interface displayed the full video alongside a timeline, with candidate time intervals pre-populated from automatic speech transcription as an initial reference. The transcription was quite accurate in locating the initiation of the recording commands, however, participants had to review and adjust these intervals to match their exact intentions. This process involved tweaking the boundaries to ensure that the selected video or photo segment corresponded precisely to the moment they intended to record. Participants were able to exclude content they didn't want to capture, ensuring that the annotations reflected their true intent at the time of recording.

For each identified moment, participants created a command entry by selecting the time interval and specifying the intended capture modality. They were also required to provide a free-text *explanation* field, describing the motivation or context behind their decision to record that particular moment (e.g., "interesting mural on the wall" or "wanted to remember this recipe").

Two distinct annotation workflows were followed, depending on the selected capture modality:

Photo Annotation. When participants selected *make a photo* as the command type, they were required to identify the specific object or scene they intended to photograph by drawing bounding boxes on the relevant video frames within the annotated interval. Each bounding box was labeled with the object and included an *expectation* field, where participants described either what they hoped to capture or communicate through the photograph or how they envisioned the photo's aesthetic qualities. The bounding boxes could be drawn across multiple frames within the same command interval, allowing participants to mark objects that appeared briefly or at varying positions. These annotated crops served as the ground truth for the photo object prediction benchmark, where models were tasked with identifying the correct object from a set of four candidate image crops extracted from the video scene.

Video Annotation. When participants selected *record a video* as the command type, they were asked to complete a structured video segment descriptor in addition to the free-text explanation. This descriptor included the start and end times of the intended video recording, a summary of the ongoing activity in free text, and three categorical attributes characterizing the nature of the moment: the moment type, selected from the categories *Information Moment*, *Visual Aesthetic Curiosity*, *Emotional Reaction*, *Social Interaction*, *Task/Goal Oriented*; the attention level, chosen from *High*, *Normal*, *Low*; and the degree of planning, selected from *Pre Planned*, *On-the-spot Plan*, *Spontaneous*. This structured data was required for all video annotation entries, ensuring that each video capture was accompanied by a detailed semantic description. The focus levels (High, Normal, Low) were

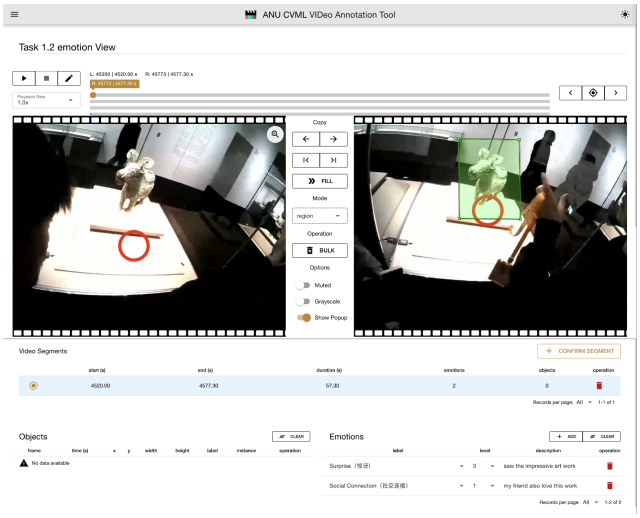


Figure 7: Annotation interface for Task1.2 (Emotion Analysis).

used to help participants better characterize their attention during the recording, with task-oriented moments typically showing higher attention and curiosity-driven moments showing lower attention.

The quality of annotations was ensured through the presence of an experimenter from the research group, who guided participants through the UI, provided clarifications when needed, and ensured that all tasks were completed properly and in the correct format.

The result of this annotation process is a fully labeled timeline for each video session. This timeline includes both the intended capture moments, with their associated semantic attributes, object localizations, and personal motivations, providing a rich, structured dataset for training and evaluating models on egocentric recording intention prediction.

A.2 TASK1.2 - EMOTION ANALYSIS

Participants annotated emotion-related moments on the Vidat platform through a retrospective review of their full egocentric recordings. To facilitate recall of fleeting affective experiences, the interface displayed the full video together with audio and gaze-overlay playback, allowing participants to inspect the surrounding context of each moment. During data collection, whenever participants noticed a salient emotional response, they sent a free-text instant message to the experimenter. These messages were not treated as labels themselves, but served as temporal anchors during retrospective annotation.

Emotion Segment Annotation. Based on the message timestamps, participants located the corresponding moments on the timeline and reviewed the surrounding footage to determine whether a noticeable emotion was present. For each identified emotional moment, they created or refined an annotation segment by selecting a time interval around the anchor timestamp. They were instructed to adjust the segment boundaries so that the annotated interval covered not only the emotional reaction itself but also the triggering event that gave rise to the emotion. In addition to the anchor-guided review, participants were also allowed to add new emotional segments while watching the full recording if they recalled other noticeable affective experiences that had not been marked during data collection.

Emotion Description and Intensity. For each segment, participants could assign one or multiple emotion labels selected from a predefined set consisting of *Joy*, *Surprise*, *Stress*, *Social Connection*, *Achievement*, *Disappointment*, *Awkward*, *Boring*, *Relaxation*, with an additional *Other* option for custom entries. They also provided a free-text *reason* field to explain what caused the emotion, describe how they felt in that moment, and optionally clarify the relative importance of multiple selected labels. In addition, participants rated the self-perceived emotion intensity on a 1–5 scale, where 1 indicated a very weak feeling and 5 indicated a very strong feeling.

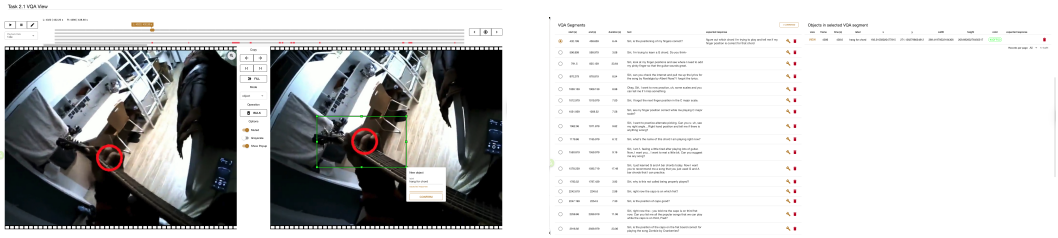


Figure 8: Screenshot for task2.1's annotation UI.

Trigger Localization. When the emotional response was clearly triggered by a visible object or person in the scene, participants were asked to draw a bounding box on the frame where that trigger could be clearly identified (e.g., if a participant felt joy upon seeing a friend, they would draw the bounding box on the frame where the friend became visible). Throughout the annotation session, an experimenter was present to explain the annotation rules, help participants refine ambiguous temporal boundaries, remind them of potentially missed emotional moments, and verify that all required fields were completed in a consistent format.

The result of this annotation process is a structured collection of emotion segments for each recording session. Each annotated entry includes the start and end timestamps of the emotional moment, one or multiple emotion labels with their corresponding intensity levels, the participant's free-text reason, and, when applicable, the bounding-box coordinates and associated frame for the visible object or person that triggered the emotion.

A.3 Task2.1 - User-Initiated Contextual Request

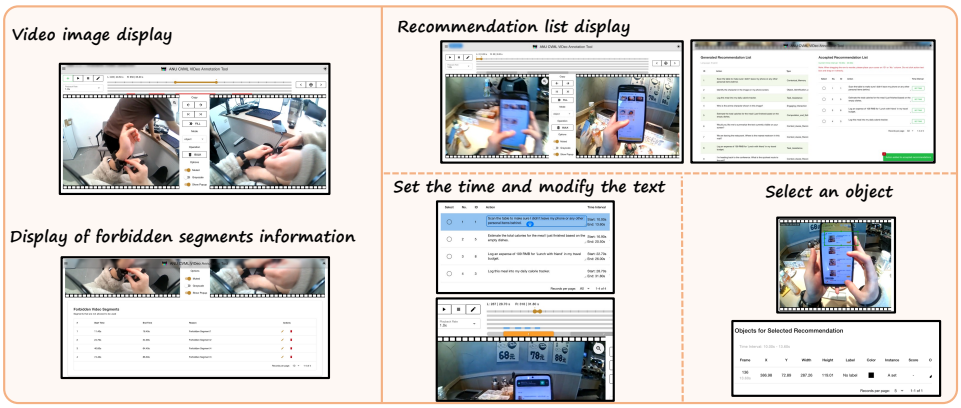
Annotation Platform. As shown in Figure 8, the interface for Task2.1 is organized around three main components. First, a timeline control panel allows annotators to navigate the full recording and define the temporal span of each query event using two frame-accurate handles. The panel supports playback at multiple speeds, keyboard-based frame navigation, and visual indicators for previously annotated VQA segments, enabling annotators to quickly review and revise existing annotations. Second, a dual-view canvas displays the video frames corresponding to the start and end handles. For visually grounded queries, annotators can draw bounding boxes around the referred objects, resize or reposition them, and assign open-vocabulary labels. Each object annotation is linked to a selected VQA segment, ensuring that visual referents are structurally associated with the corresponding spoken query. Third, two tables summarize the current annotations: the VQA segment table stores the temporal boundaries, query text, and expected response for each query event, while the object table lists the bounding boxes and object-level expected responses associated with the selected segment.

Annotation Procedure and Output. For each recording session, the experimenter helps upload the participant's egocentric video together with the automatically extracted request events generated by the preprocessing pipeline, including their interval timestamps and transcribed text. Then, the participant reviews and refines these automatically generated entries. After confirming the request content, the participant writes the requirements describing what an ideal AI assistant should provide, such as the required information type, level of detail, or preference for where to find external knowledge. If the request refers to a visible object or scene entity, the participant further annotates its visual grounding by drawing bounding box on it's corresponding frame. The label describes what the object is and how it is related to the request.

The final annotation is exported as a JSON file. Each exported file contains a list of request segments. Each segment includes the corrected start time, end time, duration, verified request transcript, segment-level expected response, and an optional list of visual referent annotations.

A.4 Task2.2 - Proactive Request Recommendations

Task2.2 annotations were collected with two complementary modules. The first module supports Proactive Timing Judgment (PTJ) by asking participants to mark intervals in which a proactive



(a) Proactive Timing Judgment (PTJ) (b) Valuable Interaction Identification (VII)

Figure 9: Task2.2 annotation platform

assistant should not interrupt them. The second module supports Valuable Interaction Identification (VII) by asking participants to review, select, edit, order, and temporally localize generated proactive recommendations.

Video availability selection module. Participants reviewed the egocentric recording with timeline-based playback and marked any interval during which they would not want to receive a proactive request and provide the reason. Each marked non-interruptible interval was visualized as a red bar on the timeline. The interface also displayed the corresponding start and end timestamps together with the participant-provided reason, allowing participants to verify and revise the annotations before export. These annotations provide the ground truth for PTJ, where the benchmark evaluates whether a model can identify contexts in which interruption is inappropriate.

Recommendation selection and reordering module. For each candidate video segment, the interface displayed the fixation-overlaid video together with a list of proactive recommendations generated by the recommendation pipeline. Participants indicated which recommendations they would accept in the observed context, optionally revised the wording to better match their intended request, reordered accepted recommendations by preference or priority, and annotated the time window in which each accepted recommendation should be presented. These annotations provide the ground truth for VII, where the benchmark evaluates whether a model can select the participant-preferred proactive request from a set of candidates. The module includes four main interface components:

1. **Video playback window.** The video player shows the egocentric clip, enabling participants to evaluate recommendations against the visual context.

2. **Recommendation list area.** The interface lists all candidate recommendations for initial screening. Participants select recommendations they would be willing to receive; selected items are added to an accepted list that can be reordered by drag-and-drop. For each accepted item, participants annotate the appropriate presentation time window, which is shown as a blue bar on the timeline to support timing verification.

3. **Recommendation editing area.** Participants can refine the wording of an accepted recommendation or rewrite it entirely when the generated text captures the right intent but not the preferred expression.

4. **Object grounding tool.** When a recommendation refers to an ambiguous object or region, participants can draw a bounding box on the video frame to identify the intended referent. The interface records the bounding-box coordinates and frame timestamp alongside the recommendation annotation.

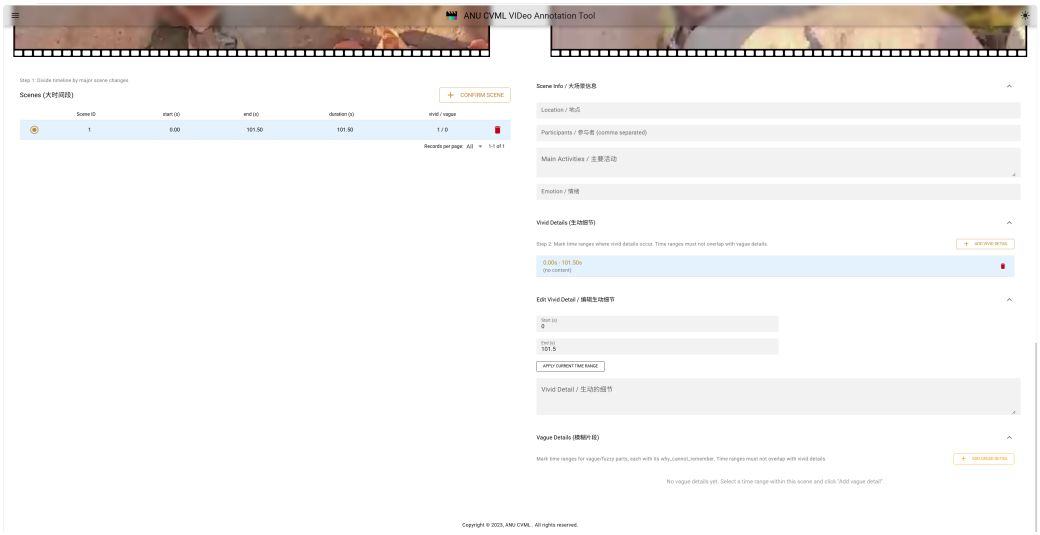


Figure 10: Task3.1 annotation platform

A.5 TASK3.1 - MEMORY RECALL PREDICTION

To ensure high-quality and consistent benchmarks, we developed a systematic pipeline to transform participants' raw episodic memories into structured, multi-level annotations.

Data Preprocessing. Following each recording session, participants provide a detailed account of the events they vividly recall, guided by a structured template that includes high-level event settings, locations, participants, primary activities, vivid details, and overall impressions. We accept accounts in both audio and text formats. These original files are processed through a standardized pipeline: audio files (e.g., .m4a) are first transcribed into text using the Xunfei or ElevenLabs API. Subsequently, a Large Language Model (LLM) is employed to decompose these vivid details into more granular, atomic segments.

Annotation. The annotation is performed on the Vidat platform. After uploading the video, experimenters execute the following systematic steps: first, they establish timestamps for each high-level event setting; second, within each event setting, they document the specific location, participants, and primary activities. Finally, experimenters determine the timestamps for each atomic vivid detail. In cases where a detail occurs multiple times or is difficult to locate within the video, we consult with the participants for clarification. The process concludes by recording the exact content for each identified vivid detail. All annotations are completed by experimenters who strictly adhere to the participants' original accounts to ensure data integrity. See Fig. 10 for details of annotation platform for MRP.

A.6 TASK3.2 - MEMORY INTENT MODELING

In-situ Labeling. During the data collection phase, participants are instructed to provide real-time verbal cues whenever they encounter an event or information they wish a "memory agent" to record. Specifically, participants trigger the labeling by saying, "Hey assistant, help me remember...", optionally followed by a brief description of the target information.

Retrospective Annotation. During the subsequent in-lab session, we employ an automated pipeline to transcribe the audio and extract the in-situ identifiers along with their corresponding timestamps into a structured JSON file. These data, along with the egocentric videos, are then integrated into the Vidat platform for refined annotation. Participants first calibrate the precise timestamps for each memory trigger. They then provide multi-dimensional metadata for each item, including the *memory content*, *motivation (reason)*, and *memory type* (e.g., Possibly Useful Information, Task Reminder, Permanent Precious Memory, Knowledge/Insights, or Custom). For object-related memories, participants are required to draw a bounding box around the target object in a representative frame. Finally, each item is categorized as either **short-term** or **long-term** based on its projected utility: information intended

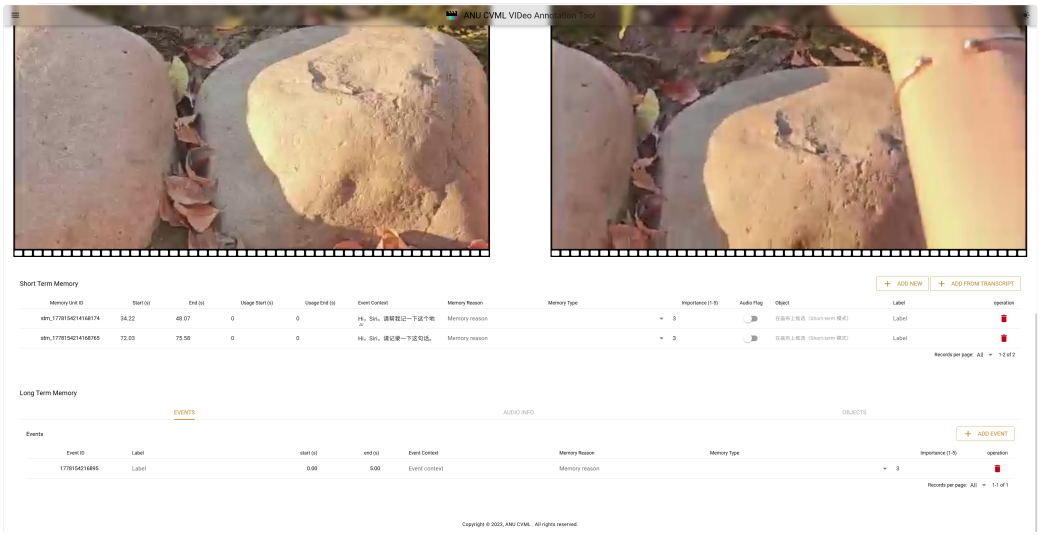


Figure 11: Task3.2 annotation platform

for use within 24 hours is classified as short-term, while all other instances are labeled as long-term. See details in Fig. 11

A.7 Task3.2 - Memory Intent Modeling

B Data Processing Algorithms for Annotation

B.1 Transcript and Command Extraction

The transcription pipeline processes continuous audio, extracts word-level transcriptions, identifies user commands directed at an AI assistant, and classifies them into categories.

Stage 1: Audio Loading & Chunking

Input: Audio file (WAV format)

The audio is split into chunks for processing. Each chunk represents a segment of the original audio.

$$\{A_0, A_1, \ldots, A_{N-1}\}, \quad N = \left\lceil \frac{\text{len}(A)}{C_{\text{ms}}} \right\rceil$$

Each chunk is defined as $A_i = \text{audio}[i \cdot C_{\text{ms}} : (i+1) \cdot C_{\text{ms}}]$.

Stage 2: Transcription Engine Selection

Three engines are supported:

- **Xunfei:** Word-level transcription with timestamps.
- **ElevenLabs:** Word-level transcription with timestamps and speaker ID.
- **GPT-4o:** Plain text transcription without timestamps.

$$T_{\text{xunfei}}(A_i) \to \{w_j = (\text{text}_j, \text{start}_j, \text{end}_j)\}_{j=1}^{M_i}$$

$$T_{\text{eleven}}(A_i) \to \{w_j = (\text{text}_j, \text{start}_j, \text{end}_j, \text{speaker}_j)\}_{j=1}^{M_i}$$

$$T_{\text{gpt}}(A_i) \to \text{text}_i$$

Stage 3: Temporal Alignment

Timestamps for each chunk are adjusted using a time offset:

$$t_{\text{offset},i} = i \cdot C_{\text{ms}}, \quad w_j^{\text{adj}} = (\text{text}_j, \text{start}_j + t_{\text{offset},i}, \text{end}_j + t_{\text{offset},i})$$

Stage 4: Full Transcription Aggregation

All chunk transcriptions are combined into a single transcript:

$$\mathcal{T} = \bigcup_{i=0}^{N-1} T(A_i) = \{w_1^{\text{total}}, \ldots, w_K^{\text{total}}\}, \quad K = \sum_{i=0}^{N-1} M_i$$

Stage 5: Command Parsing

Commands directed at the AI assistant are identified using GPT-4.1. A context window around the trigger phrase is extracted and analyzed. Retry logic is applied up to three times:

$$P : \mathcal{T} \to \{c_k\}_{k=1}^{L}$$

Stage 6: Command Classification

Parsed commands are classified as:

$$c_k = \begin{cases} \text{recording} & \text{if command matches \{“make photo”, “record video”\}} \\ \text{memory} & \text{if command matches \{“remember”\}} \\ \text{general} & \text{otherwise} \end{cases}$$

Stage 7: Output Structuring

Commands are organized by category:

$$C_{\text{recording}} = \{c : \text{type}(c) = \text{recording}\}, \quad C_{\text{memory}} = \{c : \text{type}(c) = \text{memory}\}, \quad C_{\text{general}} = \{c : \text{type}(c) = \text{general}\}$$

Each command includes:

- type: Command category
- original_text: Exact transcript text
- command: Normalized command
- explanation: User reasoning (generated by GPT, later reviewed)

B.2 Proactive Recommendation Generation Pipeline for Task2.2

The proactive recommendation pipeline converts continuous egocentric recordings into ranked interaction requests for Task2.2. Its goal is to propose interaction requests that a smart-glasses assistant could proactively surface at moments when the participant is willing to be interrupted. For each recording, it uses the egocentric video, participant-marked non-interruptible intervals, participant profile, scene-context annotations, fixation logs, and world-camera timestamps. It outputs a record per processed segment, including segment metadata, global and local summaries, fixation proportions, intermediate candidate pools, usage statistics, and final ranked proactive recommendations. The pipeline combines three sources of context: (i) the full egocentric video segment, (ii) object-level gaze fixation, and (iii) participant-specific profile and scene metadata. As summarized in Figure 12, it first gates out non-interruptible intervals, extracts fixation-aware visual context, generates global and local candidates in parallel, and reranks and deduplicates the resulting pool.

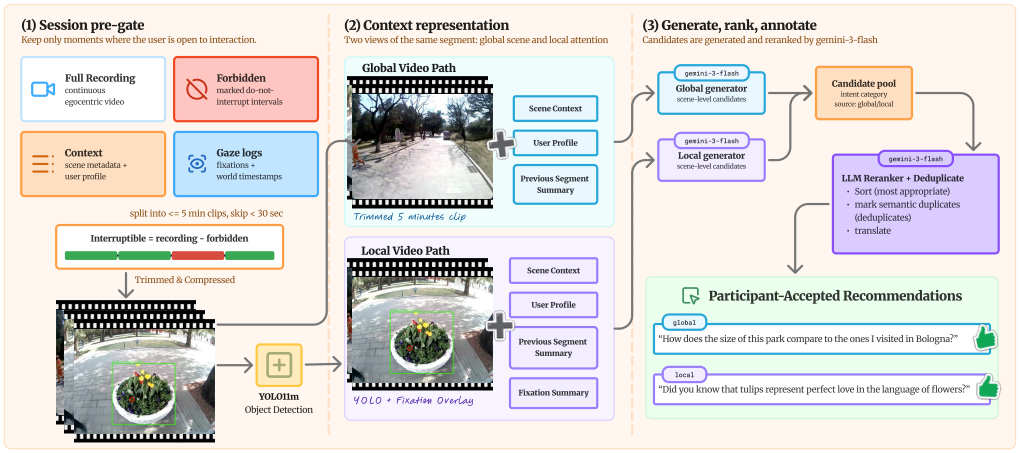


Figure 12: Overview of the Task2.2 proactive request recommendation generation pipeline. The system first removes participant-marked non-interruptible intervals, constructs five-minute interruptible segments, extracts global video context and local fixation-aware context, generates candidate recommendations with `gemini-3-flash-preview`, reranks and deduplicates them, and exports the full per-segment output for participant review.

Stage 1: Interruptibility Pre-gating and Segment Construction. The inputs are full egocentric video $\mathcal{V}$, participant-marked forbidden intervals, scene-context metadata, participant profile, fixation CSV, and world-timestamp CSV.

Participants retrospectively mark intervals in which proactive recommendations should not be shown, such as privacy-sensitive, socially inappropriate, or otherwise undesirable moments. Let the raw forbidden intervals be

$$\mathcal{B} = \{[b_i^s, b_i^e]\}_{i=1}^{M}.$$

Only the portion of each interval within the valid recording window contributes to the complement, and overlapping spans are merged into a single non-interruptible set:

$$\widetilde{\mathcal{B}} = \text{Merge}\left(\{[\max(0, b_i^s), \min(T, b_i^e)] \mid b_i^e > b_i^s\}_{i=1}^{M}\right).$$

The complementary interruptible intervals over the full recording duration T are

$$\mathcal{I} = [0, T] \setminus \bigcup_{[s,e]\in\widetilde{\mathcal{B}}} [s, e].$$

Each interruptible interval is split into segments of at most 300 seconds, with segments shorter than 30 seconds discarded. For each retained segment $v_k = [t_k^s, t_k^e]$, the original 1600×1200 recording is trimmed and compressed to height 480 pixels, 500k video bitrate, and 2 FPS. The segment inherits the scene-context annotation with the largest temporal overlap, and the previous segment summary is carried forward as temporal context.

Stage 2: Visual Object Detection. For each segment, an Ultralytics YOLO11m obeject detector (Jocher et al., 2023) returns, for each processed frame f_j, a set of detections

$$D_j = \{(o_{j,l}, b_{j,l}, p_{j,l})\}_{l=1}^{n_j},$$

where $o_{j,l}$ is the predicted object class, $b_{j,l} = (x_1, y_1, x_2, y_2)$ is its bounding box, and $p_{j,l}$ is the confidence score. These detections define the object vocabulary used by the local recommendation path and separate attended objects from broader background context.

Stage 3: Gaze–Video Synchronization. Fixation logs and processed video frames differ in timestamp space and spatial resolution. Gaze coordinates are calibrated in the original 1600×1200

world-camera coordinate system, while detection and generation operate on the compressed clip. After removing non-finite rows, the synchronizer filters both streams to the current segment:

$$\tau_k^s = \tau_0 + 10^9 t_k^s, \quad \tau_k^e = \tau_0 + 10^9 t_k^e,$$

where τ_0 is the first world-camera timestamp in nanoseconds, t_k^s and t_k^e are segment-relative times in seconds, and τ_k^s, τ_k^e are the corresponding gaze-log bounds. World-camera timestamps are indexed to match the 2 FPS processed video. Fixation coordinates are scaled to the processed resolution by

$$x' = x \cdot \frac{W_{\text{video}}}{W_{\text{orig}}}, \qquad y' = y \cdot \frac{H_{\text{video}}}{H_{\text{orig}}}.$$

Fixations that cross segment boundaries are clipped, and their durations are recomputed, yielding fixation points that can be compared directly with YOLO boxes.

Stage 4: Fixation Proportion Computation. The fixation module converts gaze behavior into an object-level attention distribution. For each clipped fixation $g_r = (x_r, y_r, \tau_r^s, \tau_r^e, d_r)$, where (x_r, y_r) is the scaled fixation point and $d_r = (\tau_r^e - \tau_r^s)/10^6$ is its duration in milliseconds, the corresponding video frame is found by timestamp lookup. If the fixation point falls inside one detected object box, the duration is credited to that object class. If multiple boxes contain the point, non-person objects are prioritized over person boxes, and the smallest box in the selected group is used. Fixations outside all detected boxes remain visible in the overlay video but are excluded from the denominator of $\rho(o)$.

The accumulated fixation duration for each object class is normalized into a percentage:

$$\rho(o) = \frac{\sum_{r:\text{hit}(g_r)=o} d_r}{\sum_{o'} \sum_{r:\text{hit}(g_r)=o'} d_r} \times 100.$$

Objects below a 5% fixation threshold are removed, and at most the top five object classes are retained. The remaining proportions are renormalized to integer percentages that sum to 100:

$$F_k = \{(o_1, \hat{\rho}_1), \ldots, (o_m, \hat{\rho}_m)\}, \quad m \leq 5.$$

The resulting fixation summary F_k is serialized as lines such as "menu: 52%" or "laptop: 31%". The pipeline also renders a fixation-overlaid clip with gaze points and fixation-hit boxes. This rendered video is used by local generators and rerankers that benefit from an explicit visualization of what the user attended to.

Stage 5: Parallel Global and Local Recommendation Generation. The core recommendation stage has two complementary generation paths.

Global path. The global generator receives the processed clip, scene context, participant profile, and previous-segment summary. It produces a short factual video summary and proactive candidates grounded in the overall scene, such as navigation help, task support, or contextual reminders.

Local path. The local generator receives the fixation-overlaid clip, fixation summary F_k, participant profile, scene context, and previous-segment summary. Its outputs are restricted to objects in F_k, capturing opportunities such as object identification, text translation, product comparison, artifact logging, or memory-related questions about attended objects.

Both generators use a structured response schema:

$$G(v_k) \to (\text{summary}, \mathcal{R}_k^{\text{global}}), \qquad L(\tilde{v}_k, F_k) \to (\text{summary}, \mathcal{R}_k^{\text{local}}),$$

where v_k is the compressed clip and $\tilde{v}_k$ is its fixation-overlaid counterpart. The final generation pipeline fixes both global and local generators to `gemini-3-flash-preview`, selected through qualitative analysis and trial runs for its price-performance trade-off on grounded, diverse, and well-structured recommendations. Structured outputs are enforced with response models and retry logic. The nine output categories are: (1) object identification and recognition; (2) translation, text, and audio recognition; (3) procedural guidance; (4) decision support; (5) contextual memory; (6) computation and estimation; (7) task assistance; (8) context-aware recommendation; and (9) engaging interaction. The prompts require concise first-person phrasing, realistic assistant-executable actions, English outputs, and strict scene grounding. The global and local paths run in parallel when fixation data are available; otherwise, the pipeline uses the global path only. Table 3 summarizes the generation and reranking prompt templates; the full executable prompts are released with the pipeline repository.

Table 3: Prompt templates used by the Task2.2 recommendation generation pipeline.

Prompt	Inputs	Core Instructions / Output
Global generator	Video clip; scene context; participant profile; previous-segment summary	Produce a factual video segment summary and proactive questions, commands, or interactions for each intent category. Recommendations must be grounded in the visible scene, personalized to the participant profile, actionable by a smart-glasses assistant, concise, first-person, English, and non-hallucinatory.
Local generator	Fixation-overlaid clip; fixation summary; scene context; participant profile; previous-segment summary	Produce a factual video segment summary and fixation-grounded recommendations. All recommendations must directly relate to objects in the fixation summary and avoid broad scene-level suggestions handled by the global path.
Reranker	Candidate pool; scene context; participant profile; previous-segment summary; fixation summary	Sort all candidates from most to least appropriate, mark semantic duplicates without dropping inputs, move ungrounded or hallucinated candidates to the bottom, translate recommendation text to Simplified Chinese, and provide a short ranking reason while preserving the original English recommendation and category fields.

Stage 6: Candidate Pool Construction. The structured outputs from both generators are flattened into a single candidate pool:

$$\mathcal{C}_k = \{(r_i, z_i, s_i)\}_{i=1}^{N},$$

where r_i is the recommendation text, z_i is the intent category, and $s_i \in \{\text{global}, \text{local}\}$ records the source path. This tag is preserved so downstream analysis can compare broad scene-level recommendations with fixation-specific recommendations.

Stage 7: LLM Reranking and Semantic Deduplication. `gemini-3-flash-preview` is also used as the LLM reranker. Given the candidate pool, scene context, participant profile, previous summary, and fixation summary, the reranker sorts candidates from most to least appropriate, checks grounding, marks semantic duplicates, translates recommendations into Simplified Chinese for Chinese-speaker participants while preserving the English source text, and provides a short ranking reason. The structured call validates that the multiset of input candidates matches the multiset of output candidates, preventing dropped, added, or rewritten candidates. Candidates marked as semantic duplicates are then removed, and the remaining recommendations receive monotonically increasing IDs.

Stage 8: Output Structuring for Annotation. For each processed segment, the full output contains segment metadata, original and fixation-overlaid clip paths or URLs, participant profile, scene context, previous summary, global and local video segment summaries, fixation proportions, intermediate candidate pools, token usage, and final ranked recommendations. These records are imported into the retrospective annotation workflow, where participants review the ranked recommendations, select those they would accept, and annotate when they should be presented. The generation pipeline therefore supplies diverse, context-grounded candidates, while participant feedback provides the ground truth for Task2.2.

C Detailed Experiment Pipeline

Each experimental session comprised two consecutive parts: (1) a 2–4 hour in-the-wild recording session with continuous wearable sensing, and (2) a 3–4 hour annotation session conducted on the same day or after a short interval. All procedures were approved by the IRB; participants provided written informed consent and received compensation at no less than local minimum wage. Participants were required to wear smart glasses without prescription eyeglasses (contact lenses permitted). Prior to recording, participants completed the `Scene context` field in the Participant Metadata spreadsheet and were instructed to imagine a smart-glasses assistant supporting photo/video capture, memory assistance, and open-domain QA, issuing all commands prefixed by in-situ identifier.

Device	Placement	Role	Data Stream
Pupil Labs Neon Glasses	Head	Egocentric video + eye gaze	RGB @ 30 fps, 2D gaze
τ-Ring	Index Finger	Physiological + behavioral sensing	HRV, PPG, IMU, temp.
Apple Watch (S8/9/11)	Dominant wrist	Physiological sensing	Accel., heart rate, IBI
BioVital S2 Wristband	Non-dominant wrist	Physiological sensing + event marker	EDA, sEMG, motion, temp., 8-ch. PPG, button-press
Samsung Galaxy S25 / S25 Ultra	Bag or Pocket	Glasses host + recording proxy	—

Table 4: Wearable hardware stack.

C.1 Wearable Hardware Stack

Table 4 summarizes the sensor suite. The Pupil Labs Neon Glasses were tethered to a Samsung phone running the Neon Companion app. The τ-Ring was worn on either hand; the BioVital S2 Wristband on the non-dominant wrist; and an Apple Watch on the dominant wrist. All devices were time-synchronized to the egocentric video via visual calibration procedures described below.

Four Samsung phones were used across sessions: Galaxy S25 (12 GB RAM, 256 GB storage; units S25-1, S25-2, S25-3) and Galaxy S25 Ultra (12 GB RAM, 512 GB storage; unit S25Ultra-1). Apple Watch Series 8, 9, and 11 were used interchangeably on the dominant wrist across the project. Data from all Apple Watch series are treated as equivalent for the core physiological signal stream (accelerometer, heart rate, IBI). The glasses draw power from the paired phone via data cable; the phone battery was verified sufficient for the full session before departure.

Evolution of the wrist-worn stack. The wrist-worn sensing stack evolved during early piloting. The first sessions used a naive self-developed wristwatch produced by an early-stage vendor. This device had no explicit start/stop button and required post-hoc temporal alignment via a PPG-based black/white boundary method: the experimentor rapidly moved the wristband across a high-contrast edge while the glasses recorded the scene; the sharp PPG signal change at each boundary crossing was later matched to the corresponding video frame. Data quality from this device proved unreliable, and the device was discarded after initial piloting. It was replaced by the BioVital S2 Wristband, which provides dense multi-modal physiological sensing, an integrated event-marker button, and auto-start on wear. The BioVital S2 simultaneously acquires seven biological modalities: EDA, sEMG, accelerometer, gyroscope, skin temperature, dynamic pulse rate and blood oxygen via an 8-channel PPG sensor, and heart rate variability (HRV).

C.2 Pre-recording Device Preparation and Calibration

The experimenter executed the following steps in strict order, recording all timestamps to the nearest second.

Pupil Labs Neon Glasses. Connected the glasses to the Samsung phone via data cable. Launched the Neon Companion app and granted camera sync permission when prompted. Verified the egocentric feed and eye-tracking overlay via the preview button; the red circle indicates the estimated gaze point. Created a participant profile and performed eye-tracking calibration by tapping the preview screen. The participant fixated a target while the experimenter dragged the gaze dot to the true fixation point until converged. Tapped record and immediately locked the phone screen to prevent accidental stoppage. Recorded the glasses start timestamp.

τ-Ring. Placed the ring in the charging case; a green light in the upper-left corner indicates active charging. The magnetic charging interface is position-sensitive and may require several attempts to seat correctly. Launched the OpenRing app, paired via `Dashboard` → `Scan Ring` → `Start Scanning`, selected the target ring from `Discovered Devices`, and tapped `Connect`. On the `Dashboard`, tapped `Update Time` then `Calibrate`. In the `Offline` tab, tapped `Get File List`; formatted storage (`Format File System`) if non-empty. Set both `Total Exercise Duration` and `Exercise Segment Duration` to 14,400 seconds (4 hours), then tapped `Start Exercise`. The ring records autonomously and stops automatically; it cannot be manually interrupted. Recorded the ring start timestamp.

BioVital S2 Wristband. The participant wore the wristband on the non-dominant wrist; logging starts automatically when worn and stops on removal. Verified the display showed `recording`. Recorded the wristband start timestamp.

Apple Watch. Worn on the dominant wrist. Launched WatchSensor-watchOS from the bottom application dock and started recording. Recorded the watch start timestamp.

Cross-device alignment. After setting up all the devices, the participant needs to clap three times while looking at their hands, separating hands immediately after each clap, and holding still for approximately one second between claps. The IMU acceleration peak at each clap was matched to the video frame showing hand contact. This procedure was repeated at the end of the session to mark the end boundary. The experimenter recorded both alignment timestamps. Post-hoc, the detected IMU peak was temporally offset to its corresponding video frame, yielding a per-device time offset to align the ring and Wristband stream to the egocentric video timeline.

C.3 In-the-Wild Recording Tasks

Participants performed their planned daily activities while wearing the full sensor stack and issued voice commands with the following minimum quotas:

Task 1: Photo / video capture (suggests $\geq$ 10 commands). Format: "Hey Assistant, take a photo" or "Hey Assistant, start recording," optionally with a description (e.g., "this sculpture is hilarious"). Purpose: capture visual moments deemed worth recording.

Task 2: Open-domain questions / commands (suggests $\geq$ 20 commands). Situationally grounded, non-trivial questions spanning definition ("Hey Assistant, What is this dish called?"), explanation ("Hey Assistant, Why is my commute longer today?"), comparison ("Hey Assistant, Which product fits my needs better?"), subjective retrieval ("Hey Assistant, Did I feed my cat this morning?"), arithmetic, assistance, translation, and objective information ("Hey Assistant, Is parking allowed here?").

Task 3: Memory assistance Format: "Hey Assistant, help me remember this" for prices, routes, phone numbers, or to-do items expected to be useful shortly after utterance.

Command marking. Each command was marked by pressing and holding the lower (marker) button on the BioVital S2 Wristband for approximately 2 seconds; a confirmation page appeared, and the participant slid the close slider to dismiss it. The upper (red) button was power only; accidental long-press would trigger shutdown and was avoided during the session.

Emotion self-report. On salient emotional events, participants sent themselves a timestamped message (e.g., "1", "surprised", "frustrated", "bored") via WeChat or any time-stamped messaging app. These timestamps guided later emotion annotation.

Scene selection guidelines. Participants were advised to: choose well-lit hours for outdoor activities; include at least one walking/moving segment; select scenarios where questions naturally arise (avoiding prolonged silent activities); ensure hands and manipulated objects were visible to the egocentric camera; and engage actively in multi-person interactions when applicable.

C.4 Post-recording Shutdown and Data Export

Shutdown order: (1) IMU clap alignment (end boundary), (2) wristband removed (auto-stops), (3) Apple Watch stopped, (4) glasses stopped, (5) participant completed `User profile` in the Participant Metadata spreadsheet.

Glasses export. Raw data was exported from the Samsung phone to the PC via USB. Cloud-processed data was uploaded via the Neon Companion app to Pupil Cloud with the network proxy active. Both raw and processed data were saved to the participant's raw data folder.

Video compression. The exported video was compressed to produce a concatenated video with audio and a muted variant for annotation.

Ring export. Via the OpenRing app, the current session was downloaded to the phone and transferred to the PC. The resulting binary log ($\sim$10 MB nominal; $\sim$300 B indicates low battery; 0 B indicates firmware failure) was saved to the participant's raw data folder.

Wristband export. The BioVital S2 was connected to the PC via USB and restarted; it mounted as a USB drive. Files were copied to a local import folder and processed through the BiovitalLab app. To create a project for the participant, import the matching record, and export to CSV format. The exported data was uploaded to the participant's raw data blob container.

Watch export. Apple Watch data were synced to the paired iPhone and uploaded to the participant's raw data folder.

C.5 Annotation Phase

The annotation phase consisted of one mandatory same-day task and several subsequent tasks scheduled after a short delay to balance memory fidelity with participant availability.

Same-day tasks.

- **Task 2.2.1: Do-not-disturb segments.** Using Vidat, the participant stated personal interruption criteria then annotated all segments where they would not want the assistant to intervene. Exported as `forbidden_segments_Px_Name.json`, serving as prerequisite input for the recommendation-generation algorithm.
- **Task 1.2: Emotion annotation.** The participant identified segments eliciting salient emotional reactions, drew bounding boxes around triggering objects when visible, selected emotion label and intensity, and optionally provided free-text explanation. Exported as `annotations_task1.2_Px_Name.json`.

Between-session preprocessing.

1. **Audio extraction.** Audio was extracted from the compressed video for downstream transcription.
2. **Speech transcription.** A Whisper-based pipeline was applied to the extracted audio to produce word-level transcripts and parse verbal commands. This yielded five JSON files stored in the transcripts folder: `all_transcripts`, `all_commands`, `general` (open-domain questions), `memory` (memory commands), and `recording` (photo/video commands).
3. **Recommendation generation** (for Task 2.2.2). The recommendation pipeline consumed the egocentric video, participant profile, scene context, fixation logs, and do-not-disturb segments to generate ranked proactive recommendations for each interruptible segment. The output was a combined JSON file imported into Vidat for participant review.

Subsequent annotation tasks (with-in one week).

- **Task 3.1.1: Oral recall.** Following each recording session, participants provided a detailed retrospective account of events they vividly recalled. They were audio-recorded while freely narrating everything they remembered from the recording period. A structured template guided the narration: high-level event settings (e.g., restaurant, commute, gym), locations, participants, primary activities, vivid details (salient objects, surprising moments, sensory impressions), forgotten items or fuzzy memories, overall emotional valence, and a self-rated memory completeness score (0–5 integer). Accounts were accepted in both audio (e.g., `.m4a`) and text formats; most participants chose to narrate orally. Audio files were transcribed into text using the Xunfei or ElevenLabs API. The resulting transcript, whether originally text or transcribed from audio, was then processed through an LLM pipeline that

decomposed free-form recollections into atomic, timestamp-granular detail segments for downstream annotation.

- **Task 1.1: Moment-of-record prediction.** The audio-derived command JSON was loaded into Vidat. For each detected photo command: the participant drew a bounding box on the target object, labeled the category, and filled `action description`, `type`, `attention`, and `intention` (planned beforehand / on-site / spontaneous). For video commands: selected the target segment. Missed commands were added manually with `+COMMAND`.
- **Task 2.1: VQA.** The open-domain command JSON was loaded into Vidat. The participant verified each detected question: deleted erroneous or duplicate entries, corrected transcription errors, and wrote the Expected Response. Bounding boxes were drawn for deictic references ("this," "that," "it").
- **Task 2.2.2: Recommendation selection.** The algorithm-generated recommendation JSON was uploaded into Vidat. The participant selected useful information, excluded unwanted suggestions, ranked by preference, and drew bounding boxes for ambiguous object references.
- **Task 3.1.2: Memory recall annotation.** The transcribed oral-recall output was loaded into Vidat. Experimenters established timestamps for each high-level event setting, documented location, participants, and primary activities within each setting, and localized each atomic vivid detail to its precise temporal boundaries. The exact content of each detail was recorded per the participant's original account.
- **Task 3.2: Memory annotation.** The memory command JSON was loaded into Vidat. Short-term: verified each "help me remember" utterance, edited text, and completed `reason`, `type`, anticipated `usage`, `importance`, and `usage_time`; drew bounding boxes and `Role` field for visually grounded memories; marked audio segments for auditory memories. Long-term: the participant voluntarily added any additional information worth remembering with free-text justification.

D More on Related Works

D.1 Egocentric Datasets and Benchmarks

D.1.1 Egocentric Dataset Characteristics

Existing large-scale datasets, such as Ego4D (Grauman et al., 2022) and EgoExo4D (Grauman et al., 2024), have significantly advanced egocentric vision. However, as shown in Table 1, these datasets typically follow a *Staged* capture protocol and rely on *external* annotators, which often limits the data to observable physical actions.

In contrast, **EgoIntrospect** introduces a **subject-chosen** capture protocol where participants record continuous sessions (up to 240 minutes) reflecting their natural daily lives. Distinct from prior works like EgoLife (Yang et al., 2025), our dataset is the first to integrate **Physiological Signals** (e.g., PPG, GSR) alongside video and gaze. Crucially, we employ **Self-annotation**, addressing the intrinsic ambiguity of "mind-reading" in egocentric vision. By obtaining ground truth directly from the subjects, we provide authentic labels for internal states that remain inaccessible to external observers.

D.1.2 Egocentric VQA Benchmarks

As illustrated in Table 2, existing benchmarks like EgoSchema (Mangalam et al., 2023) and HourVideo (Chandrasegaran et al., 2024) primarily evaluate external reasoning tasks such as spatial-temporal grounding (G), forecasting (F), and planning (P). Their questions are largely formulated by researchers or LLMs.

EgoIntrospectQA expands the scope of egocentric understanding to a multi-dimensional task suite (**A&I&M**). Our benchmark stands out in three aspects:

1. **Subject-Sourced Intelligence:** Both questions and ground-truth are generated by the **subjects**, ensuring that tasks reflect genuine human cognitive processes rather than external heuristic guesses.
2. **Internal-External Fusion:** We challenge MLLMs to infer internal states, including **Affective Experience (A)**, **Interaction Intent (I)**, and **Cognitive Memory (M)**. This allows for a more holistic assessment of agents in understanding both "what is happening" and "why the user feels/acts so."
3. **Format Diversity:** Moving beyond standard MCQ aDnd OE formats, we incorporate **Multiple Formats** (*), including Binary Classification (BC) and Multi-class (MC) Classification (e.g., 9-way classification), providing a more granular evaluation of model performance across diverse task requirements.

D.2 Affective Experience, Request Intent and Cognitive Memory

Affective Experience. A key challenge for revisiting daily experiences with smart glass (Hodges et al., 2006; Yao et al., 2016) is how to align recorded moments with what is meaningful to the user (Le et al., 2016). Existing work suggests that such moments can be indicated by both users' emotional states (Ståhl et al., 2009; Sas et al., 2013) and attentional behavior (Chang et al., 2021; Li et al., 2021). Related HCI research has also explored affect sensing using diverse signals, including facial expressions (Masai et al., 2017; Li et al., 2024; Kwon et al., 2021), eye images (Wu et al., 2020), and physiological signals (Jammot et al., 2025). Recent works have begun to study emotion understanding in egocentric settings, mb by jointly considering first-person visual context and wearable signals (Zhao et al., 2022; Cheng et al., 2024b; Hu et al., 2025; Jammot et al., 2025). Leveraging this line of work, our study specifically probes whether MLLMs can address tasks related to affective experiences.

Request Intent. Several HCI studies aim to resolve under-specified requests using natural multimodal context, including deictic language, gaze, pointing, and dialogue history (Wang et al., 2024; Lee et al., 2024; Yan et al., 2024; Peng et al., 2025; Sun et al., 2025). More proactive settings further reduce explicit input (Lee et al., 2025): the system infers user's unspoken request at a cued moment, a problem related to visual question generation (Mostafazadeh et al., 2016) and egocentric intent modeling (Pan et al., 2026; Veerabadran et al., 2025) but centered on user-specific needs. One step further, proactive recommendation requires systems to infer not only what assistance may be useful, but also when it can be delivered without disruption. (Lee et al., 2025; Cai et al., 2025; Zhang et al., 2025a). *EgoIntrospect* supports under-specified request understanding, proactive request prediction, and request recommendation by collecting user-centric annotations.

Memory. Continuous recording of egocentric videos have long been used for lifelogging, offering an external memory pool to help users recall past experiences (Hodges et al., 2006; Gelonch et al., 2019; Sellen et al., 2007). A central challenge lies in how to identify memorable moments (Niforatos et al., 2017; Sellen & Whittaker, 2010). These efforts can facilitate AI assistants for specific applications, such as support for cognitive impairment (Ait Gacem et al., 2019; Li et al., 2019), way-finding (Qiu et al., 2024), and other situated memory tasks. Recent wearable memory agents further integrate egocentric sensing with LLMs or AI models to retrieve, summarize, and proactively present memory cues in real time (Zulfikar et al., 2024; Haddad et al., 2025; Datta et al., 2022; Yuan et al., 2025). Instead of addressing researcher-designed memory tasks as in prior work, our study focuses on both what users are likely to remember themselves and the lifespan of memory retrieval.

E Dataset and Benchmark Statistic

E.1 Capture Devices

To capture a comprehensive representation of user states and environmental interactions, we employed a multi-device wearable platform capable of recording diverse data modalities. The hardware configuration is designed to synchronize first-person visual perspectives with fine-grained physiological and motion signals.

As summarized in Table 5, our primary visual and gaze data are captured using the **Pupil Labs Neon Glasses**, which provide high-resolution egocentric video alongside precise eye-tracking metrics. To monitor internal physiological responses, participants wore a **BioVital S2 Wristband** and a

Smart Ring, providing redundant yet complementary signals such as photoplethysmography (PPG), electrodermal activity (EDA), and skin temperature. Furthermore, an **Apple Watch** was utilized to collect high-frequency inertial measurements (IMU) and environmental context (e.g., GPS, altitude, and ambient volume).

This multi-modal setup allows for the joint analysis of external stimuli and internal states. All devices were time-synchronized to a common clock to ensure temporal alignment across video, motion, and physiological streams, providing a rich basis for understanding human intent and affective status in-the-wild.

Device	Recorded Data Modalities
Pupil Labs Neon Glasses (Baumann & Dierkes, 2023)	First-person RGB video, gaze (position, azimuth), pupil data, IMU (accelerometer, gyroscope, quaternion), audio, recording events
BioVital S2 Wristband	PPG (green 4-channel, red, infrared), heart rate, respiratory rate, EDA/GSR, SpO_2, skin temperature, device temperature, ambient temperature and humidity, atmospheric pressure, IMU (accelerometer, gyroscope), 3-axis magnetometer
Apple Watch	IMU (accelerometer, gyroscope, quaternion, gravity, magnetic field) at $\sim$30 Hz; heart rate, step count, altitude, activity, GPS, volume at $\sim$1 Hz; SpO_2, respiratory rate, body temperature, HRV at $\sim$1/60 Hz;
Smart Ring (Tang et al., 2025)	PPG (green, red, infrared), IMU (accelerometer, gyroscope), temperature (3 channels)

Table 5: Devices and Recorded Data Modalities

E.2 Demographics

Our dataset comprises a diverse group of 60 participants, with ages ranging from 15 to 48 years ($\mu = 23.63$, $Mdn = 23$). In terms of gender distribution, the cohort is relatively balanced, consisting of 33 females (55%) and 27 males (45%). The participants are internationally diverse, representing 14 different nationalities. While Chinese participants form the majority ($n = 41$), the dataset also includes representation from Indonesia ($n = 3$), Spain ($n = 2$), Italy ($n = 2$), Nepal ($n = 2$), and one participant each from Bulgaria, France, Mexico, Kazakhstan, Ghana, and the Arabian region. This variety in age, gender, and cultural background ensures a broad spectrum of perspectives and behaviors within the collected data.

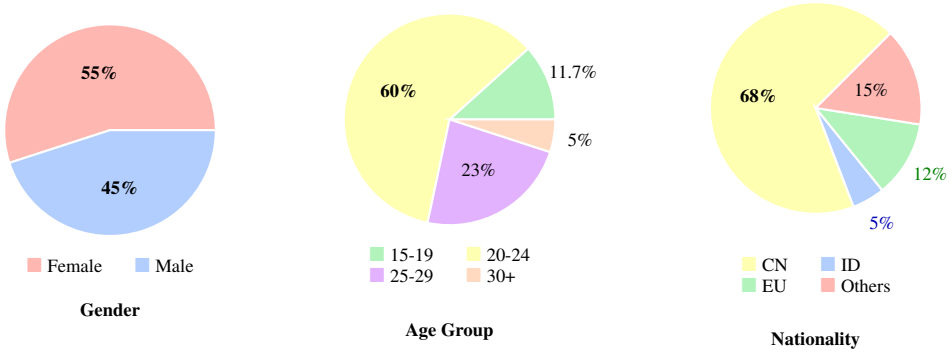


Figure 13: Demographic statistics of the 60 participants.

E.3 Scenario Coverage

Rather than recording isolated tasks, each participant contributed a single, continuous long-form video session lasting between 2 to 4 hours. This continuous recording protocol was designed to capture the natural transitions between different environments and daily activities. Specifically, each

long-form session covers 2 to 4 distinct scenario categories, resulting in a total of 192 annotated activity segments across 14 diverse categories.

The distribution of these scenarios is detailed in Table 6 and illustrated in Fig. 14. Our dataset encompasses a broad spectrum of human activities, ranging from structured professional tasks to spontaneous social interactions and high-dynamic physical movements.

Table 6: Statistics of Scenario Categories and Activity Examples.

Scenario Category	**Activity Examples**	**Segments**
Social Interaction	Multi-person conversations, group activities, caregiving	33
Eating	Dining in canteens, restaurants, cafés, and bakeries	27
Leisure & Special Events	Park walking, art exhibitions, dog walking, live performances	22
Shopping	Malls, department stores, supermarkets, and open markets	16
Sports	Light strength training, jogging, skateboarding	14
Transportation/Navigation	Navigating public transit, airports, and guide environments	12
Entertainment	Arcade centers, board game rooms, watching sports matches	10
Travelling & Visiting	Cultural/historical sites and natural attractions	10
Personal Living (Home)	Cooking, household tidying, and deep cleaning	9
Tools, Assembly, Manufacture	Pottery, LEGO building, woodworking, and metalworking	9
Outdoors & Gardening	Cycling, hiking, and gardening	9
Business/Service Processing	Parcel pickup, service counter interactions	8
Creative Arts & Performance	Instrument practice, drawing, calligraphy, and knitting	7
Work and Study	Laboratory work, clinical shadowing, and electrical operations	6
Total		**192**

The data collection process adhered to strict ethical and privacy guidelines. For scenarios involving high privacy sensitivity, such as *Personal Living* or *Social Interaction*, explicit consent for public data sharing was obtained from both the participants and their identifiable social partners. To ensure data quality and relevance for egocentric vision research, we intentionally avoided purely stationary desk-based activities (e.g., simple reading or computer work) and strictly prohibited the recording of copyrighted content. This collection strategy ensures a rich and ethically-compliant representation of dynamic, real-world human behavior.

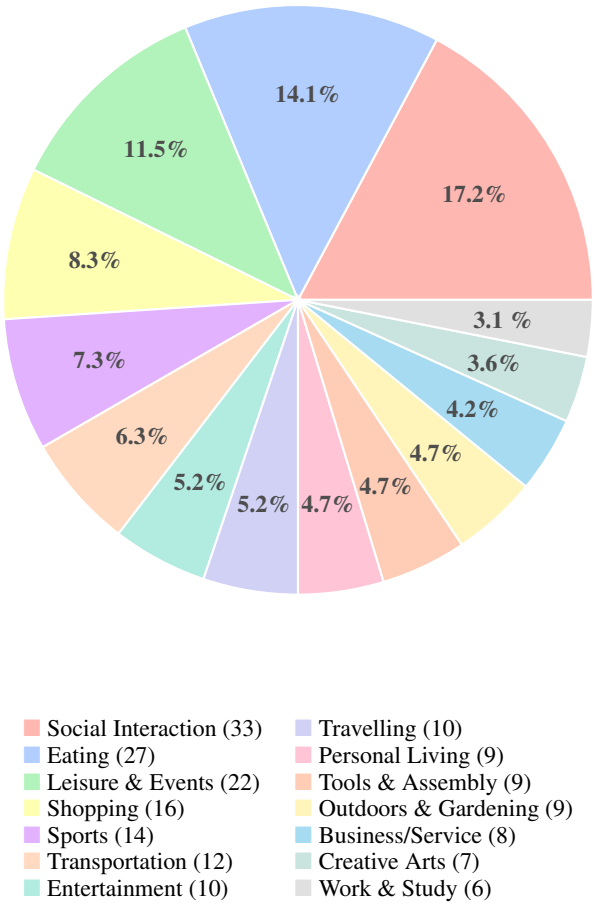
E.4 Dataset Split

To evaluate model generalization while supporting future model development, we partition the dataset based on **unique participants**. This user-independent splitting strategy ensures that the test participants are never seen during model development, reducing the risk of participant-level data leakage.

Our current dataset contains 60 participants. We reserve a fixed set of 15 participants as the **Test Set** for final benchmark evaluation and cross-model comparison. The remaining 45 participants are used for development and are split into a **Training Set** and a **Validation Set** following an 8:2 ratio, resulting in 36 training participants and 9 validation participants. The training set is intended for model weight optimization, while the validation set is used for hyperparameter tuning, prompt selection, and model selection. All test participants are strictly excluded from the development phase.

As data collection is ongoing, newly collected participants will be assigned to the training and validation sets following the same 8:2 ratio, while the fixed test set will remain unchanged to provide a stable benchmark for future comparison. The current 60-participant split is summarized in Table 7.

Figure 14: Distribution of scenario categories across the 192 annotated activity segments. Each segment was extracted from 60 long-form egocentric video sessions, covering 14 diverse real-world contexts.



Distribution of 14 collection scenarios

Table 7: Dataset partition based on the current 60 participants.

Split	Participants	Percentage	Role
Training	36	60%	Model Development
Validation	9	15%	Prompt/Model Selection
Test	15	25%	Final Evaluation
Total	**60**	**100%**	–

E.5 Benchmark Video Rendering

All video-based benchmark tasks use a shared fixation-overlay renderer to standardize model inputs. For each task instance, the source egocentric recording is trimmed to the task-specific temporal window, compressed for inference, and augmented with visual cues that expose the state of the eye-tracking signal. As shown in Figure 15, frames with an active fixation display a red circular marker at the fixation location. If gaze tracking is available but no fixation is detected at that frame, the video displays `[no fixation]`; if the gaze log indicates that the glasses are not worn, or that gaze tracking is unavailable, it displays `[not worn]`. This representation separates three visually similar cases—attended gaze, missing fixation, and unavailable gaze—so models can reason about the scene without treating absent fixation markers as implicit evidence of user attention. For evaluations that use overlaid video, the system prompt defines the red marker and both status labels before presenting the task-specific question.



(a) Valid fixation (b) No fixation (c) Not worn

Figure 15: Examples of gaze-state overlays in video-based benchmark inputs. The shared renderer exposes whether a frame contains an active fixation marker, has valid gaze tracking but no detected fixation, or corresponds to a period in which the glasses are not worn or gaze tracking is unavailable.

F Dataset Open-Source

F.1 Open-Source Plan

We plan to release our dataset in two complementary forms: benchmark data and the full dataset. The benchmark data will be publicly released to support direct reproduction of the benchmark results reported in this paper. This release will include all processed benchmark instances used in our evaluation, including the request-centered video clips, benchmark JSON files, task prompts, candidate answers, ground-truth labels, and evaluation scripts. The released benchmark data will be organized to work directly with our benchmark codebase, enabling future researchers to evaluate new models under the same task definitions and input settings.

In addition to the benchmark data, we will release the full dataset through a controlled-access process. The full dataset includes the continuous egocentric recordings, time-aligned multimodal signals, and participants' original retrospective annotations across all task categories. Since these recordings contain first-person video and may include bystanders, private spaces, screens, speech, and other sensitive contextual information, access to the full dataset will require a data-use agreement and an approval process. Similar to the access model used by large-scale egocentric video datasets such as Ego4D (Grauman et al., 2022), approved users will receive access credentials for downloading the controlled-access data.

Our current collection contains 60 participants. The benchmark results reported in this paper are computed on the 15-participant test set, while the remaining 45-participant training set will be released as part of the controlled-access full dataset. We plan to make the full dataset available around the conference release time. Since this project is designed as an ongoing data collection effort, we will continue to expand the controlled-access release with newly collected and processed data over time. Through this staged release plan, we aim to make the benchmark fully reproducible while preserving appropriate privacy protection for the broader first-person dataset.

F.2 De-identification Process

We will adopt a de-identification process similar to Ego4D (Grauman et al., 2022), combining automatic detection with manual review before releasing any public-access data. Since almost all of our recordings include outdoor or semi-public scenarios, the videos may contain bystanders who did not provide explicit consent for identity disclosure. Therefore, all public benchmark clips will be reviewed and de-identified before release.

The primary de-identification target is human faces. We will first apply an automatic face-detection and blurring pipeline to all video clips. After automatic processing, trained reviewers will manually inspect the results to correct missed detections and remove false positives. This manual verification step is necessary because first-person videos often contain challenging cases, such as motion blur, partial faces, side views, reflections, occlusions, or small faces in the background. Beyond faces, we will also manually review the videos for other sensitive information. In our data, we observed that participants may occasionally capture highly private content during daily activities, such as password entry, payment interfaces, addresses, or other personally identifiable information. These regions will be manually marked and blurred before public release. If a clip contains sensitive content that cannot be safely de-identified without substantially damaging the task content, it will be excluded from the public-access benchmark release.

We will also remove direct identifiers from released metadata, filenames, transcripts, and annotation files. Participant names and session identifiers will be replaced with randomized IDs, and any personally identifying text in transcripts or participant-written annotations will be removed or generalized. This process ensures that all publicly released benchmark data undergoes de-identification while preserving the information necessary for model evaluation.

G Benchmark Task1.1 - Moment of Recording

The benchmark for "Moment of Recording" evaluates the ability of AI models to predict and interpret user-initiated recording actions in egocentric video data.

G.1 Benchmark Question Definition and Curation Pipeline.

This task includes two primary question types:

1. **The Recording Test**, which assesses whether a given video segment contains content that would prompt a user to initiate recording.
2. **The Photo Test**, which identifies the specific object or scene that a user chose to photograph from a set of options.

These questions probe the model's ability to infer user intent, visual appeal, and attentional focus from a first-person perspective.

The curation pipeline starts with an annotated egocentric video dataset, where human annotators identify moments when users initiate recording actions (e.g., taking photos). Each annotation includes the temporal boundaries of the recording event, a textual description of the user's intent, and the objects of interest in the scene.

For the **Recording Test**, positive windows are constructed by extending 30 seconds before and after each user-initiated recording. This provides additional context to help the model understand the user's intent while ensuring that the recording windows are over one minute in length, as many participants tend to select recording moments that are just a few seconds long. Negative windows are sampled from non-overlapping video segments, and we include a greater number of negative windows to better simulate real-life scenarios, where users most often do not record anything. This also helps to avoid bias by ensuring that the model is not overly conditioned to expect a recording in every segment. The input consists of the scene context, requirements for the task, and the positive or negative segmented video. The output is a yes or no answer for whether recording should happen or not.

For the **Photo Test**, the pipeline also includes video segments where a photo was explicitly taken, being extended in a similar way with 30 seconds before and after each photo moment. The ground-truth target is determined by the annotated objects in the scene. To create a multiple-choice question,

three additional distractor objects are selected randomly from a 10-second window around the exact moment the photo was taken. This encourages the model to evaluate which of the objects in the scene was most likely to have been selected for the photo. We also ensure that distractor objects are not repetitive across test groups, fostering a more challenging task for the model. The input consists of the scene context, requirements for the task, and the segmented video around the photo action, and 4 photos. The output is a choice of A/B/C/D corresponding ot the photo that the user should have recorded.

It is important to note that the video segments given to the models do not include audio in order to prevent the model from understanding if the recording was initiated through the voice command of the user.

G.2 Benchmark Statistic.

The performance of models in both the **Recording** and **Photo Tests** is primarily evaluated using accuracy, which measures the proportion of correct predictions to the total number of predictions.

Recording Test: For the Recording Test, the model is tasked with predicting whether a given video segment contains content that would prompt a user to initiate recording. The expected output is a binary classification: "yes" if the content in the segment is likely to prompt recording, and "no" if it is not. The goal is to assess how accurately the model identifies moments in which a user might initiate recording based on the visual content.

Photo Test: For the Photo Test, the model is asked to predict the object or scene that a user would choose to photograph from a set of options. The expected output is a multiple-choice selection of the correct object from a set of four candidate image crops extracted from the video. The goal is to evaluate the model's accuracy in identifying the object the user intends to photograph.

G.3 Benchmark Results.

Dataset overview. The recording test (Subtask 1a) comprises **744** fixed-duration video windows drawn from 15 participants: 253 positive windows (34.0%), each centred ± 30 s around an annotated recording command (`record a video` or `make a photo`), and 491 negative windows (66.0%), randomly sampled from non-recording intervals and matched in duration to their positive counterparts. These window counts are fixed and model-independent; all evaluated systems face the same 744 instances. Positive windows decompose into 189 originating from `make a photo` commands and 63 from `record a video` commands; the latter carry additional annotations for segment type (Information Moment, Visual Aesthetic Curiousity, Task/Goal Oriented, Social Interaction, Emotional Reaction), attention level (Low / Normal / High), and intention (Spontaneous / On-the-spot Plan / Pre Planned), enabling a fine-grained diagnostic analysis. The photo-identification test (Subtask 1b) contains **181** four-choice multiple-choice questions, one per `make a photo` command, where the model must identify the intended capture target among four candidate object crops extracted from the same video segment. Random chance performance is 25%; the correct-answer distribution across positions is balanced (A: 53, B: 38, C: 50, D: 40), ruling out positional bias. Human annotators achieve **77%** accuracy on the recording test and **83%** on the photo test.

G.3.1 Recording Test: GPT-4o vs. InternVL3.5-8B-Instruct Analysis

We select GPT-4o and InternVL3.5-8B-Instruct for a detailed error analysis because InternVL3.5-8B-Instruct is the top-performing model overall (68.41% accuracy), narrowly surpassing GPT-4o (68.15%) while being a compact 8B open-source model. The comparison reveals that comparable aggregate accuracy can mask meaningful differences in where and why each model succeeds.

Shared conservative bias, different precision-recall balance. Table 9 shows that both models adopt a conservative strategy, preferring to abstain from predicting a recording moment rather than risk a false alarm. Both achieve specificity above 81%, while recall remains below 46%. GPT-4o has a slight edge in recall (45.6% vs. 42.7%), recovering six more true positives (114 vs. 108), whereas InternVL3.5 is marginally more conservative with higher specificity (81.7% vs. 81.5%) and fewer false alarms (90 vs. 98). The result is a near-identical positive-class F1 (50.3% vs. 47.9%), confirming

Table 8: Task 1.1 benchmark results across all evaluated models. Recording test: 744 windows (binary yes/no). Photo test: 181 four-choice MCQ instances (random baseline 25%). All values in %.

Models	Recording Test		Photo Test	
	Acc.	Macro-F1	Acc.	Macro-F1
Closed-source models				
Gemini-3-flash-preview	58.24	57.99	**60.92**	60.95
Gemini-3.1-pro-preview	67.70	63.65	54.80	54.36
GPT-4o	68.15	63.53	24.86	25.98
Kimi-K2.5	63.58	62.83	29.28	21.83
Kimi-K2.5(thinking mode)	53.22	53.12	21.54	18.17
Qwen3.5-397B-A17B	56.32	57.34	24.31	20.18
Qwen3.5-Omni-Plus	56.85	57.11	26.52	20.42
Qwen3.6-Plus	56.59	56.66	24.86	21.26
Open-source models				
InternVL2.5-8B	57.66	57.34	42.54	40.14
InternVL3-8B	65.46	62.37	46.41	44.26
InternVL3.5-8B-Instruct	**68.41**	62.62	45.30	42.45
InternVL3.5-8B-Thinking	58.47	57.37	43.09	43.18
MiniCPM-V-4.5	48.32	46.91	24.58	17.65
Qwen2.5-VL-7B-Instruct	66.62	41.84	48.60	49.59
Qwen3-Omni-30B-A3B-Instruct	52.49	52.70	60.89	59.86
Qwen3-Omni-30B-A3B-Thinking	50.07	49.72	54.75	54.47
Qwen3-VL-8B-Instruct	57.74	57.74	52.51	52.07
Qwen3-VL-8B-Thinking	40.51	35.81	58.66	58.31
LongVU-Qwen2-7B	65.99	39.76	22.65	12.11
LongVA-7B-DPO	48.25	48.23	45.86	45.72
LLaVA-OneVision-Qwen2-7B	58.60	57.53	28.73	11.56
LLaVA-Video-7B-Qwen2	45.03	42.40	29.28	23.43
Human performance	77.00	65.00	83.00	91.00

that despite the open-source size advantage, neither model has found a substantially better operating point for this task.

Table 9: Recording test — positive class diagnostic metrics for GPT-4o and InternVL3.5-8B-Instruct (744 windows; 253 positive, 491 negative).

Model	**Acc.**	**Prec.**	**Recall**	**Spec.**	**$F1_{pos}$**	**TP / FN**
GPT-4o	68.1	56.2	45.6	81.5	50.3	114 / 136
InternVL3.5-8B-Instruct	68.4	54.5	42.7	81.7	47.9	108 / 145

Command-kind gap. Table 10 shows that both models follow the same pattern: recall on `make a photo` windows is roughly twice that on `record a video` windows. This consistent gap across models suggests that photo-command moments carry a visually clearer, more localised trigger — a salient foreground object highlighted by the gaze fixation — whereas video-recording moments are often motivated by broader temporal or ambient context that a short clip alone does not reliably convey. InternVL3.5 trails GPT-4o slightly on both command kinds (49.7% vs. 51.3% on photo; 22.2% vs. 27.0% on video), indicating GPT-4o retains a narrow advantage in detecting these positive moments despite the overall accuracy parity.

Table 10: Recording test — recall (%) on positive windows by command kind.

Command kind	n	**GPT-4o**	**InternVL3.5**
`make a photo`	189	51.3	49.7
`record a video`	63	27.0	22.2

Complementary strengths across segment types and intentions. Despite their overall similarity, the two models show complementary strengths when broken down by annotation category (Table 11). GPT-4o leads on Information Moment (42.9% vs. 23.8%), performing well when a legible object or informative text is centered in the field of view, but nearly fails on Task/Goal Oriented (7.1%) and

Visual Aesthetic Curiosity (14.3%), types where recording intent has no single immediately detectable visual anchor. InternVL3.5 exhibits the reverse strength: it is best on Task/Goal Oriented (28.6%) and more balanced across types (21–29%), though uniformly lower than GPT-4o on Information Moment. Neither model detects Emotional Reaction moments (both 0%), confirming that affectively motivated recording is currently beyond the reach of both systems.

The intention dimension reveals **GPT-4o's most critical failure**: On-the-spot Plan moments yield only 3.3% recall (one out of 30), while Spontaneous recordings reach 58.8%. This gap implies the model depends almost entirely on immediate visual triggers and is blind to short-horizon deliberation. InternVL3.5 avoids this extreme collapse (On-the-spot Plan: 16.7%; Spontaneous: 23.5%; Pre Planned: 36.4%), though its uniformly lower recall means it does not offer a practical advantage on any individual intention class.

Table 11: Recording test — recall (%) by segment type and intention (`record a video` windows only, n as labelled; G4o = GPT-4o, IVL = InternVL3.5-8B-Instruct).

Segment type				Intention			
Type	n	**G4o**	**IVL**	**Intention**	n	**G4o**	**IVL**
Information Moment	21	42.9	23.8	Spontaneous	17	58.8	23.5
Task/Goal Oriented	14	7.1	28.6	Pre Planned	11	27.3	36.4
Vis. Aesthetic Curiousity	14	14.3	21.4	On-the-spot Plan	30	3.3	16.7
Social Interaction	6	33.3	16.7				
Emotional Reaction	3	0.0	0.0				

Participant-level variance. Both models show extreme participant-level inconsistency with largely overlapping failure patterns. GPT-4o achieves 95.0% recall for P2 yet 0% for P28, P32, P44, and P5. InternVL3.5 improves on several of GPT-4o's zero-recall participants (P13: 87.5%, P33: 81.0%, P52: 66.7%) but introduces new failures of its own (P49: 0%, P45: 8.3%, P5: 11.1%) and notably still fails entirely on P28. This large inter-participant variance ($\sigma_{\text{recall}} > 0.30$ for both models) indicates that participant-specific scene characteristics — recording context, environment, and individual behaviour patterns — dominate over cross-participant model capability, and that current VLMs lack the user-adaptive grounding required for consistent proactive recording.

G.3.2 Photo Test: Insights and Discussion

Subtask 1b asks the model to watch the egocentric video clip and select, from four candidate object crops, the specific object the user intended to photograph. Unlike the binary recording test, this task demands fine-grained understanding of where the user's attention is directed and what, among visually competing objects, they considered capture-worthy.

Near-chance performance of most models. Table 8 reveals that the majority of closed-source models cluster at or below the **25% random baseline**. GPT-4o scores 24.86%, and Kimi-K2.5, Qwen3.5-397B, Qwen3.5-Omni-Plus, and Qwen3.6-Plus all fall in the 24–29% range. This near-chance performance, combined with GPT-4o producing 17 invalid responses out of 181 (9.4%), suggests these models cannot reliably connect gaze-indicated attention to a specific object, or default to positional guessing when fine-grained visual grounding fails.

Reversed model ranking. The photo test induces a markedly different model ranking than the recording test. Gemini-3-flash-preview, the weakest closed-source model on recording (58.24%), becomes the strongest overall on photo identification (60.92%), matched only by Qwen3-Omni-30B-Instruct (60.89%) and approaching the human ceiling of 83%. Conversely, GPT-4o, the top performer on recording, drops to essentially random performance on the photo test. This reversal suggests the two subtasks draw on complementary capabilities. Recording prediction benefits from holistic scene-level saliency assessment, while photo identification requires fine-grained, object-level visual grounding conditioned on inferred user intent.

Open-source models are competitive. On the photo test, multiple open-source models substantially outperform most closed-source entries. The InternVL family shows consistent improvement with

architecture version: InternVL2.5-8B (42.54%), InternVL3-8B (46.41%), InternVL3.5-8B (45.30%), all well above GPT-4o (24.86%). Qwen3-VL-8B-Instruct (52.51%), Qwen2.5-VL-7B-Instruct (48.60%), and Qwen3-VL-8B-Thinking (58.66%) further demonstrate that 7–8B open models can rival or exceed mid-range closed-source systems on this task. Thinking variants of Qwen3-VL and Qwen3-Omni consistently outperform their Instruct counterparts on the photo test (e.g., Qwen3-VL-Thinking 58.66% vs. Instruct 52.51%), consistent with the expectation that explicit chain-of-thought reasoning benefits the object-identification step.

Persistent human gap. Despite Gemini-3-flash reaching 60.92%, a gap of over **22 percentage points** to human performance (83%) remains. The difficulty appears concentrated in cases where the target object is identified not by visual prominence alone but by the user's momentary intent, encoded in subtle gaze patterns, proximity, and contextual familiarity—factors that current VLMs cannot reliably infer from a short egocentric clip.

G.3.3 Ablation Study

Table 12: Task 1.1 ablation over input conditioning for representative closed- and open-weight models (same decoding as Table 8). Scores are percentages ($\times 100$). ICL denotes a one-shot in-context example in the benchmark prompt from the same participant, to act as their past interaction history. **Bold** marks the best score in each numeric column within the Closed-source/API and Open-source/local blocks.

Model	Inputs	Recording		Photo	
		acc	macro f1	acc	macro f1
Closed-source/API models					
kimi-k2.5	video	44.22	51.21	**32.59**	**23.52**
	video + gaze	63.57	62.83	29.28	21.83
	video + gaze + ICL	**70.00**	**66.16**	20.48	20.30
Open-source/local models					
Qwen3-VL-8B-Instruct	video	52.01	51.85	**55.42**	**55.15**
	video + gaze	**57.73**	**57.73**	52.51	52.07
	video + gaze + ICL	52.01	51.18	52.97	52.40

We conduct two ablation conditions—gaze removal and in-context learning (ICL)—on both subtasks using Kimi-K2.5 and Qwen3-VL-8B-Instruct as representative models. In the no-gaze condition, the fixation overlay is omitted, and models receive the raw video without any gaze indicator. In the ICL condition, a single task-relevant demonstration (question–answer pair) is prepended to the prompt; samples without a pre-assigned example are excluded, leaving 730 recording windows and 166 photo samples. Table 13 summarizes all conditions.

Table 13: Ablation results for Kimi-K2.5 and Qwen3-VL-8B-Instruct (accuracy / macro-F1, %). Baseline = standard gaze overlay, no ICL.

	Recording test		**Photo test**	
Condition	**Kimi**	**Qwen3-VL**	**Kimi**	**Qwen3-VL**
Baseline (w/ gaze)	63.57 / 62.83	57.73 / 57.73	29.28 / 21.83	52.51 / 52.07
No gaze	44.22 / 51.21	52.01 / 51.85	32.59 / 23.52	55.42 / 55.15
ICL (one-shot)	70.00 / 66.16	52.01 / 51.18	20.48 / 20.30	52.97 / 52.40

Gaze is a critical signal for the recording test. Removing the gaze overlay causes Kimi-K2.5 to drop **19.4 points in accuracy** (63.57% to 44.22%). Qwen3-VL-8B-Instruct is less sensitive, losing 5.7 points (57.73% to 52.02%), suggesting it relies more on holistic video content and less on the fixation signal. Overall, the gaze overlay provides a substantial grounding benefit for the recording task, particularly for models that attend closely to the attentional cue.

Gaze slightly hurts photo identification. Counterintuitively, removing gaze modestly improves photo-test accuracy for both models: Kimi gains 3.3 points (29.28% to 32.59%) and Qwen3-VL gains

2.9 points (52.51% to 55.42%). A likely explanation is that the circular fixation marker introduces visual clutter that interferes with fine-grained object matching against the four candidate crops, whereas the raw frame provides a cleaner signal for visual comparison. This finding suggests that the optimal gaze representation may differ between the two subtasks.

ICL helps Kimi on recording but hurts on the photo test. One-shot ICL raises Kimi's recording accuracy by **6.4 points** (63.57% to 70.00%) and improves macro-F1 by 3.3 points, indicating that a concrete example of the yes/no judgement meaningfully guides Kimi's decision boundary. For Qwen3-VL, ICL provides no benefit on recording (57.73% to 52.02%), suggesting the model's baseline instruction-following is already well-calibrated for this task, and the example may introduce a confounding bias. On the photo test, ICL substantially hurts Kimi (29.28% to 20.48%, a drop of **8.8 points**), potentially because the single-shot example biases the model toward a particular answer format or object category that misaligns with most test instances. Qwen3-VL remains stable on the photo test under ICL (52.51% to 52.97%), showing no sensitivity to the demonstration in either direction.

H BENCHMARK TASK1.2 - EMOTION ANALYSIS

H.1 BENCHMARK QUESTION DEFINITION AND CURATION PIPELINE.

We construct the Task 1.2 benchmark from participant-annotated emotion segments and instantiate three question types that probe affect understanding at progressively increasing levels of difficulty: emotion-label prediction (EI-label), emotion-label prediction with reason grounding (EI-label+reason), and relative emotion-intensity comparison (EIR).

EI-label. EI-label evaluates whether a model can identify the participant's emotion category from a single egocentric video segment. Let $\mathcal{E} = \{e_1, \ldots, e_9\}$ denote the predefined emotion taxonomy, consisting of *Joy*, *Surprise*, *Stress*, *Social Connection*, *Achievement*, *Disappointment*, *Awkward*, *Boring*, and *Relaxation*. For an annotated segment with ground-truth label e_i^*, we define the candidate set as $\mathcal{C}_i = \mathcal{E}$ and require the model to select one option $\hat{y}_i \in \mathcal{C}_i$ given the multimodal context $\mathcal{X}_i$. Although the original annotation protocol allows an *Other* field, we adopt a fixed-label benchmark design and retain only segments that can be assigned to one of these nine categories.

EI-label+reason. EI-label+reason extends EI-label by requiring the model to jointly predict the participant's emotion and the reason underlying that emotion. For the same segment, let the correct option be $c_i^+ = (e_i^*, r_i^*)$, where e_i^* is the participant-annotated emotion label and r_i^* is the corresponding reason description. We then construct a nine-way candidate set $\mathcal{C}_i = \{c_i^+, c_{i,1}^-, \ldots, c_{i,8}^-\}$, where each distractor $c_{i,j}^- = (e_j, r_{i,j}^-)$ pairs an alternative emotion category with a Gemini-3.1-pro-generated reason conditioned on the video content, the participant's annotated reason, and the target emotion category. The model is required to select one option $\hat{y}_i \in \mathcal{C}_i$. This formulation is more demanding than category prediction alone, as it requires the model to align the observed scene with both the reported affect and its attributed cause.

EIR. Emotion Intensity Recognition evaluates whether a model can compare the relative strength of emotional experiences rather than merely identify their category. For a target participant and emotion category, we construct a four-choice candidate set $\mathcal{C}_i = \{c_{i,1}, c_{i,2}, c_{i,3}, c_{i,4}\}$, where each candidate $c_{i,j} = (\mathcal{X}_{i,j}, e_{i,j}^*, r_{i,j}^*)$ contains the available multimodal context, the annotated emotion label, and the corresponding reason description for one clip. The model is then required to identify either the strongest or the weakest emotional experience among the four candidates, with the ground-truth targets defined by the participant-rated intensity scores:

$$y_i^{\max} = \arg\max_{j \in \{1,\ldots,4\}} e_{i,j}^{I,*}, \qquad y_i^{\min} = \arg\min_{j \in \{1,\ldots,4\}} e_{i,j}^{I,*}.$$

When fewer than four valid clips are available for a target participant-emotion group, the remaining slots are filled with clips from the same participant that do not belong to the target emotion and do not share the same reason description, thereby preserving within-user comparison while reducing shortcut cues from duplicated causes.

H.2 BENCHMARK STATISTIC.

Task 1.2 contains three benchmark variants constructed from 15 participants. The two emotion-identification variants, EI-label and EI-label+reason, share the same 281 benchmark instances and differ only in the candidate answer format, while EIR contains 88 four-choice intensity-comparison questions. To ensure consistent temporal context, all benchmark clips are uniformly cropped to 180 seconds. The EI benchmark spans all nine predefined emotion categories, with a moderately imbalanced label distribution: Joy (98), Surprise (64), Stress (34), Social Connection (28), Relaxation (17), Awkward (11), Disappointment (11), Achievement (9), and Boring (9). EIR is balanced across the two question variants, containing 44 highest-intensity questions and 44 lowest-intensity questions.

Table 14: Task1.2 benchmark statistics.

Question Type	# Instances	# Participants	Answer Space	Chance
EI-label	281	15	9 labels	11.11
EI-label+reason	281	15	9 emotion-reason pairs	11.11
EIR	88	15	4 candidates	25.00

H.3 BENCHMARK RESULTS.

H.3.1 BASELINE RESULTS

Evaluation protocol. We evaluate the same set of models across all three benchmark variants. For the main benchmark setting, EI-label and EI-label+reason are evaluated under the full multimodal input condition with video, gaze, transcript, and physiological signals, while EIR is evaluated under video, gaze, and transcript inputs. Human performance is measured on a separate subset comprising roughly 10% of the benchmark data using the same input formats and answer constraints as the model evaluation. We report accuracy for all three variants and additionally report macro-F1 to account for label imbalance and option-level bias in the multiple-choice outputs.

Table 15: Task 1.2 baseline benchmark results. Scores are percentages ($\times 100$).

Models	EI-label		EI-label+reason		EIR	
	acc	macro f1	acc	macro f1	acc	macro f1
Closed-source models						
Gemini-3-flash-preview	**62.28**	61.92	62.63	60.21	**55.68**	**51.34**
Gemini-3.1-pro-preview	58.71	59.84	**66.90**	**65.37**	54.55	48.92
gpt-4o	59.79	58.66	63.35	62.47	50.00	46.87
kimi-k2.5	53.38	52.11	54.09	51.83	39.77	41.26
kimi-k2.5(thinking mode)	54.40	54.92	53.74	52.95	47.73	44.18
qwen3.5-397b-a17b	51.60	50.28	52.67	49.96	44.32	42.07
qwen3.5-omni-plus	54.09	53.41	54.44	53.18	46.60	43.95
Open-source models						
InternVL2_5-8B	27.41	26.58	29.97	28.91	**43.18**	36.27
InternVL3-8B	22.42	21.36	27.47	26.84	34.09	33.18
InternVL3_5-8B-Instruct	25.98	24.73	35.68	34.92	39.77	35.64
InternVL3_5-8B-Thinking	29.54	28.11	32.47	31.08	36.36	34.05
MiniCPM-V-4_5	14.95	13.42	38.18	37.25	38.64	34.72
Qwen2.5-VL-7B-Instruct	38.08	36.47	31.39	30.14	29.55	32.46
Qwen3-Omni-30B-A3B-Instruct	30.25	29.06	37.47	36.55	37.50	35.11
Qwen3-Omni-30B-A3B-Thinking	36.30	35.48	35.68	34.87	38.64	36.02
Qwen3-VL-8B-Instruct	30.25	28.94	31.75	30.66	42.05	**37.58**
Qwen3-VL-8B-Thinking	37.37	36.02	33.54	32.41	32.95	33.47
LongVU_Qwen2_7B	39.50	37.26	**43.06**	**41.72**	31.82	30.84
LongVA-7B-DPO	**42.00**	**40.33**	40.21	38.96	32.95	32.27
llava-onevision-qwen2-7b-ov-hf	19.22	17.88	24.97	23.41	40.91	34.93
LLaVA-Video-7B-Qwen2	28.47	27.16	28.83	27.95	35.22	31.76
human performance	78.57	74.80	89.29	86.10	60.00	55.20

Table 15 shows a consistent closed-source advantage across all three Task 1.2 variants. On EI-label, Gemini-3-flash-preview achieves the best score (62.28 accuracy, 61.92 macro-F1), while the strongest open-source result is LongVA-7B-DPO (42.00, 40.33). On EI-label+reason, Gemini-3.1-pro-preview is strongest (66.90, 65.37), whereas the best open-source score is LongVU_Qwen2_7B (43.06,

41.72). The enlarged gap in EI-label+reason indicates that jointly modeling emotion and explanation remains substantially harder than label-only prediction for open-weight models.

For EIR, Gemini-3-flash-preview remains the top performer among closed-source models (55.68, 51.34). The best open-source EIR accuracy is obtained by InternVL2_5-8B (43.18), while the best open-source EIR macro-F1 is achieved by Qwen3-VL-8B-Instruct (37.58). Compared with EI tasks, EIR macro-F1 is generally lower, reflecting the additional difficulty of relative intensity comparison over multiple clips. Overall, Task 1.2 suggests a persistent capability gap in user-centric affect reasoning, especially when moving from emotion labeling to reason-grounded and intensity-comparison settings.

H.3.2 Ablation Results

Table 16 reports ablation results for representative closed-source and open-source models. For EI-label and EI-label+reason, we compare four input settings: *video*, *video+gaze*, *video+gaze+transcript*, and *video+gaze+transcript+physiology*. For EIR, we compare *video*, *video+gaze*, *video+gaze+transcript*, and *video+gaze+transcript+ICL*.

Table 16: Task 1.2 ablation over input conditioning. Scores are percentages ($\times 100$).

Model	Inputs	EI-label		EI-label+reason		EIR	
		acc	macro f1	acc	macro f1	acc	macro f1
Closed-source/API models							
kimi-k2.5	video	53.38	52.11	54.09	51.83	39.77	41.26
kimi-k2.5	video + gaze	52.67	50.94	53.38	50.72	38.64	39.88
kimi-k2.5	video + gaze + transcript	**56.58**	**55.31**	**58.01**	**56.27**	43.18	43.62
kimi-k2.5	EI: V+G+T+P / EIR: V+G+T+ICL	56.23	54.68	57.65	55.94	**44.32**	**44.05**
Open-source/local models							
Qwen3-VL-8B-Instruct	video	30.25	28.94	31.67	30.66	42.05	37.58
Qwen3-VL-8B-Instruct	video + gaze	31.32	29.37	31.32	30.11	40.91	36.92
Qwen3-VL-8B-Instruct	video + gaze + transcript	35.59	33.08	**36.65**	**35.24**	**47.73**	**40.17**
Qwen3-VL-8B-Instruct	EI: V+G+T+P / EIR: V+G+T+ICL	**36.30**	**33.42**	36.30	34.96	46.59	39.48

The ablation results show that transcript is the most consistently useful additional signal for both representative models, while gaze-only input can be neutral or slightly harmful. For kimi-k2.5, adding gaze to video reduces EI-label ($53.38/52.11 \rightarrow 52.67/50.94$) and EI-label+reason ($54.09/51.83 \rightarrow 53.38/50.72$), whereas adding transcript yields the largest gains (EI-label $56.58/55.31$, EI-label+reason $58.01/56.27$). The final EI setting with physiology does not further improve over the transcript setting ($56.23/54.68$ and $57.65/55.94$), indicating limited incremental benefit from physiology in this configuration.

Qwen3-VL-8B-Instruct shows a similar but not identical pattern. Gaze slightly improves EI-label ($30.25/28.94 \rightarrow 31.32/29.37$) but slightly hurts EI-label+reason and EIR ($31.67/30.66 \rightarrow 31.32/30.11$, $42.05/37.58 \rightarrow 40.91/36.92$). Transcript provides the largest gain for EI-label+reason and EIR ($36.65/35.24$ and $47.73/40.17$), while the final EI setting with physiology further improves EI-label to $36.30/33.42$. By contrast, the same final row slightly underperforms transcript-only on EI-label+reason and EIR ($36.30/34.96$, $46.59/39.48$), indicating that physiology and ICL provide task-dependent gains rather than uniformly improving all subtasks.

I Benchmark Task2.1 - User-Initiated Contextual Request

I.1 Benchmark Question Definition and Curation Pipeline.

We constructed the benchmark for User-Initiated Contextual Request from the participant-verified request annotations described above. Each annotated request segment contains the corrected temporal boundary, the verified spoken request u_i^*, the participant-provided expected response requirement ρ_i^*, and optional visual referent annotations. Based on these annotations, we constructed two question types: *Tool Selection (TS)*, which evaluates whether the model can identify the external capabilities

needed to fulfill a request, and *Request Prediction (RP)*, which evaluates whether the model can infer the user's intended request from the surrounding egocentric context.

I.1.1 TOOL SELECTION (TS)

Stage1: Initial tool requirement annotation. For each participant-verified request segment, we used an LLM to infer which external tools would be needed to satisfy the request. The input to this step included the spoken request u_i^*, the expected response requirement ρ_i^*, the segment timestamps, and any associated visual referent annotations. The LLM produced two tool lists: *required tools*, which are necessary to answer the request, and *helpful tools*, which may improve the response but are not strictly required. For each selected tool, the LLM also generated example tool calls, including function names and query arguments.

Stage2: Tool name normalization and tool-universe construction. Since each request was processed independently, semantically equivalent tools could be assigned different surface names, such as `map_search`, `maps_api`, or `location_search`. We therefore aggregated all tool names across participants and used an LLM to propose a normalization mapping from raw tool names to canonical tool names. This mapping was then manually reviewed by the researchers to correct two types of errors: over-merging, where functionally different tools were mapped to the same canonical name, and under-merging, where equivalent tools remained separated.

After normalization, we rebuilt the final tool universe $\mathcal{T}$ using the canonical tool names, consisting of 33 tools in total. For each tool, we summarized its usage examples and generated a short natural-language description. We then updated each participant's tool annotations by replacing raw tool names with canonical names. When a tool name changed during normalization, we regenerated its corresponding example calls so that the function names and query arguments remained consistent with the canonical tool semantics.

Stage3: Video clip, transcripts, and benchmark instance construction. For each request segment, we extracted an egocentric video clip spanning 30 seconds before the request onset to 30 seconds after the request offset. We overlaid gaze fixations as red circles and added an "Asking Question" label during the time interval in which the participant was speaking the request. For participants with multiple recording sessions, we first processed and concatenated the session videos with gaze overlays, then extracted request-centered clips from the merged video.

Each *TS* instance presents the model with the request-centered video clip and a complete list of available tools. The candidate set consists of all canonical tools in $\mathcal{T}$ plus a "No tool needed" option. Unlike a four-option multiple-choice question, *TS* is formulated as a multi-select task: the model must output all tools needed to fulfill the request. We store the normalized required-tool list and helpful-tool list as ground truth. To support models with different input capabilities, we construct two prompt versions: one using the original user question text, and another that additionally includes an automatic transcript of the full video clip. We further transcribed the full request-centered clip, including both the spoken request and the audio context, using automatic speech recognition: Xunfei LFASR (iFLYTEK, 2026) for Chinese recordings and ElevenLabs (ElevenLabs) for English recordings.

Correctness Criterion for Tool Selection. Because each TS instance contains both *required* and *helpful* tool annotations, we use a hierarchical correctness criterion rather than a simple exact-match rule. For each instance i, let R_i denote the set of required tools, H_i denote the set of helpful-but-not-required tools, $\hat{P}_i$ denote the set of tools predicted by the model, and n denote the special option *No tool needed*. We define the strict correctness indicator as:

$$\mathrm{Correct}(i) = \begin{cases} 1, & \text{if } R_i \neq \emptyset \text{ and } R_i \subseteq \hat{P}_i \subseteq R_i \cup H_i, \\ 1, & \text{if } R_i = \emptyset, H_i \neq \emptyset, \text{ and } \left(\hat{P}_i = \{n\} \text{ or } \emptyset \neq \hat{P}_i \subseteq H_i\right), \\ 1, & \text{if } R_i = \emptyset, H_i = \emptyset, \text{ and } \hat{P}_i = \{n\}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the model must select all required tools and may additionally select helpful tools, but any irrelevant tool makes the answer incorrect.

To further characterize model behavior beyond strict accuracy, we report two auxiliary metrics for *TS*. *Required Recall* measures whether the model identifies all tools that are strictly necessary for

Question Type	# Instances	Key Statistics
Tool Selection (TS)	298	33 tools; 96 no-tool cases; 120 multi-tool cases
Request Prediction (RP)	289	4 options per instance; all clips muted during the query interval

Table 17: Summary statistics of the Task 2.1 benchmark.

fulfilling the request. It is computed only on instances with non-empty required-tool annotations and does not penalize the model for selecting additional helpful or irrelevant tools. This metric captures whether the model recognizes the core external capabilities needed by the request.

We also report *Distractor Rate*, which measures how often the model selects tools outside the annotated acceptable set. For each instance, we treat the union of required and helpful tools as the set of acceptable tools, while *No tool needed* is treated as a special option rather than a real tool. A prediction is counted as containing a distractor if it includes any real tool that is neither required nor helpful. This metric reflects the model's tendency to over-call irrelevant tools.

I.1.2 Request Prediction (RP)

Stage1: Partially muted video clip construction. RP evaluates whether the model can infer what the participant intended to ask, rather than simply recognizing the spoken words. We therefore used the same request-centered video construction as TS, including the ± 30 second temporal context, gaze fixation overlays, and the "Asking Question" label. However, for RP we muted the audio during the request interval $[t_0, t_1]$, so that models with audio understanding could not directly hear the answer. Audio outside the request interval was preserved, since surrounding environmental sounds or conversations may provide useful context.

Stage2: Candidate request generation. For each muted request clip, we used Gemini to generate a four-option candidate set. The model was given the muted video together with the participant annotation, including the original request u_i^*, the expected response requirement ρ_i^*, and any visual referent annotations. Gemini then generated one *correct request*, which reformulates the participant's actual request and expected response requirement into a complete request description, and three *wrong requests*, which are plausible in the same egocentric context but do not match the participant's true intent. The original request text was used only to help generate high-quality options and is not exposed in the final benchmark prompt.

Stage3: Benchmark instance construction. We randomly shuffled the correct request and three distractor requests using a fixed seed, assigned them to options A–D. We generated transcripts for the muted clips using automatic speech recognition. Because the request interval itself was muted, the transcript only captures the surrounding audio context rather than the participant's spoken request. Each *RP* instance asks the model to select the option that best matches the participant's actual request. The user prompt contains only the four candidate requests, and does not include the original spoken request. As with *TS*, we provide two benchmark versions: one without transcript and one with a transcript block containing only the surrounding audio context.

I.2 Benchmark Statistics.

Task 2.1 contains 587 benchmark instances from 16 participants, including 298 *Tool Selection (TS)* instances and 289 *Request Prediction (RP)* instances. Each instance is paired with a request-centered egocentric video clip, with a median clip duration of 66.55 seconds and a median spoken-query duration of 6.55 seconds. All records passed our basic quality checks, with no missing fields, invalid fields, or duplicated record keys.

I.3 Benchmark Results.

Overall, as shown in Table 18, closed-source/API models generally perform better than open-source/local models on Task 2.1, especially for *Request Prediction*. For *Tool Selection*, the best strict accuracy is achieved by Gemini-3.1-pro-preview among closed-source models and InternVL3.5-8B-Instruct among open-source models. However, high strict accuracy does not always align with high required recall: some models identify many required tools but also select many irrelevant tools,

Table 18: Benchmark Results for Request Intent - Task 2.1

Models	Tool Selection			Request Prediction	
	acc	required recall	distractor rate	acc	macro-F1
Closed-source models					
Gemini-3-flash-preview	34.46	**72.33**	62.50	**65.31**	**63.53**
Gemini-3.1-pro-preview	**52.70**	60.13	**38.18**		
gpt-4o	41.81	53.55	49.13	39.14	38.04
kimi-k2.5	43.96	64.15	45.97	48.10	46.45
kimi-k2.5(thinking mode)	50.34	59.12	39.60	49.48	47.16
qwen3.5-397b-a17b	50.67	71.70	43.29	50.17	48.26
qwen3.5-omni-plus	45.64	68.55	46.31	59.86	57.42
Open-source models					
InternVL2_5-8B	32.89	39.62	53.69	35.64	34.58
InternVL3-8B	27.52	57.86	67.11	35.99	34.69
InternVL3_5-8B-Instruct	**51.34**	60.38	39.93	38.06	36.92
InternVL3_5-8B-Thinking	40.27	59.75	48.66	40.14	39.37
MiniCPM-V-4_5	21.81	64.46	67.22	41.87	40.14
Qwen2.5-VL-7B-Instruct	24.83	59.12	69.46	39.79	37.58
Qwen3-Omni-30B-A3B-Instruct	45.30	54.09	43.29	**42.91**	**40.63**
Qwen3-Omni-30B-A3B-Thinking	40.27	49.06	48.66	42.21	40.26
Qwen3-VL-8B-Instruct	26.85	**67.92**	70.47	40.83	39.22
Qwen3-VL-8B-Thinking	39.93	49.37	48.31	41.87	39.95
LongVU_Qwen2_7B	36.91	17.61	32.89	28.72	28.53
LongVA-7B-DPO	37.58	23.27	**30.41**	35.64	34.28
llava-onevision-qwen2-7b-ov-hf	6.04	61.19	89.54	34.60	32.70
LLaVA-Video-7B-Qwen2	7.05	32.69	90.78	35.29	34.11
human performance	66.67	73.33	26.67	79.31	76.79

Table 19: Task 2.1 ablation over input conditioning for representative closed-source/API and open-source/local models.

Model	Inputs	Tool Selection			Request Prediction	
		Acc.	Required Recall	Distractor Rate	Acc.	Macro-F1
Closed-source/API models						
kimi-k2.5	video	51.68	59.12	36.58	43.25	42.02
	video + gaze	50.00	59.12	38.93	45.33	44.48
	video + gaze + transcript	43.96	64.15	45.97	48.10	46.45
	video + gaze + transcript + ICL	–	–	–	50.36	48.49
Open-source/local models						
Qwen3-VL-8B-Instruct	video	27.18	57.86	66.11	30.45	29.61
	video + gaze	27.52	59.75	66.11	30.10	29.05
	video + gaze + transcript	26.85	67.92	70.47	40.83	39.22
	video + gaze + transcript + ICL	–	–	–	39.78	38.77

leading to a high distractor rate. This reflects the difficulty of the task, where models must not only recognize the necessary external capabilities but also avoid over-calling unrelated tools.

For *Request Prediction*, Gemini-3-flash-preview achieves the best performance, substantially outperforming both open-source models and other API models. This suggests that inferring the user's intended request from egocentric context remains challenging, especially when the spoken request itself is removed from the audio. Compared with *Tool Selection*, this task benefits more clearly from additional contextual signals, since the model must rely on visual context, gaze, and surrounding audio cues to infer what the user was trying to ask.

The ablation results in Table 19 further show different effects of input conditioning across the two question types. For *Tool Selection*, adding gaze and transcript does not consistently improve performance and may even reduce strict accuracy or increase distractor rate. One possible reason is that the explicit user request is already provided in the prompt, so additional contextual signals may introduce noise or encourage models to over-select tools. In contrast, *Request Prediction* consistently benefits from richer context. For kimi-k2.5, performance improves from video-only to video+gaze, then further improves with transcript and in-context examples. A similar pattern appears for Qwen3-VL-8B-Instruct, where adding transcript leads to a large gain in both accuracy and macro-F1. These results indicate that gaze and surrounding audio are especially useful when the model must infer the latent request rather than reason from an explicit user query.

J Benchmark Task2.2 - Proactive Request Recommendations

J.1 Benchmark Question Definition and Curation Pipeline.

As introduced in Section 4.3.2, Task2.2 evaluates proactive assistance in egocentric streams from two complementary perspectives: **when** should the assistant interrupt the user and **what** proactive interaction would be valuable and helpful if an intervention is allowed. We instantiate these perspectives as two benchmark question types.

Proactive Timing Judgment (PTJ). PTJ is a binary classification task that asks whether the user is interruptible during a video segment. Each sample contains a fixation-overlaid egocentric video clip, the participant profile, and, when available, the audio transcript. The model must output exactly one label:

$$y \in \{\texttt{True}, \texttt{False}\},$$

where `True` means that a proactive recommendation can be shown, and `False` means that the segment should not be interrupted.

To construct PTJ, we convert each participant's retrospective interruptibility annotation into a temporally grounded binary supervision signal. For a recording of duration T, participant-marked forbidden intervals are first clipped to $[0, T]$; after merging overlapping spans, the resulting non-interruptible set is denoted by $\widetilde{\mathcal{B}}$. Its complement,

$$\mathcal{A} = [0, T] \setminus \bigcup_{[s,e] \in \widetilde{\mathcal{B}}} [s, e],$$

defines the candidate interruptible intervals. To keep model inputs within a comparable temporal scale, intervals from both $\widetilde{\mathcal{B}}$ and $\mathcal{A}$ are partitioned into clips of at most 300 seconds, while very short fragments below 15 seconds are discarded. Clips derived from $\widetilde{\mathcal{B}}$ receive the `False` label, while clips derived from $\mathcal{A}$ receive the `True` label. Each retained clip is then materialized as a fixation-overlaid egocentric clip and paired with the participant profile, transcript, temporal metadata, and the binary prompt. Participant-provided explanations for forbidden intervals are preserved as metadata but are not exposed as target answers. Thus, the PTJ input is an egocentric video clip, and the output is one binary interruptibility label.

Valuable Interaction Identification (VII). VII is a four-choice multiple-choice task that asks which candidate recommendation should be proactively offered in a given segment. Each sample contains a fixation-overlaid video clip, participant profile, scene context, previous-segment summary, optional transcript, and four recommendation options:

$$\mathcal{C}_i = \{c_A, c_B, c_C, c_D\}.$$

The model outputs a single option key $\hat{y}_i \in \{A, B, C, D\}$. The correct option is a participant-accepted recommendation; the other three options are distractors sampled from recommendations generated for the same recording segment but not accepted by the participant.

To construct VII, we start from the merged proactive-recommendation annotations, which combine the generated candidate pool with participants' retrospective acceptance decisions. For segment k, let $\mathcal{P}_k$ be its generated recommendation pool and $\mathcal{A}_k$ be the participant-accepted subset. Each accepted recommendation $a_i \in \mathcal{A}_k$ becomes one benchmark instance and serves as the positive option. When a participant edited the accepted recommendation, the edited text is used as the ground-truth option; Chinese edits are translated into English to keep the benchmark prompt language consistent. The negative options pool excludes all accepted IDs from the same segment,

$$\mathcal{N}_{k,i} = \{r \in \mathcal{P}_k \mid \mathrm{id}(r) \notin \mathrm{id}(\mathcal{A}_k)\},$$

from which three negative options are sampled without replacement using a fixed seed. The four options are then shuffled with the same seeded random generator to avoid positional bias, and the resulting correct option key is recorded. The visual input is centered on the participant-annotated presentation window, for an accepted interval $[u_i^s, u_i^e]$, the clip interval is

$$[\max(u_i^s - 60, 0),\ u_i^e],$$

where $[u_i^s, u_i^e]$ is first replaced by the full source segment interval when the annotation is marked as $[-1, -1]$. Each benchmark instance retains the accepted recommendation, negative option IDs, option strings, correct key, scene context, previous-segment summary, transcript field, and fixation-overlaid clip path. Thus, the VII input is an egocentric video clip plus four candidate proactive requests, and the output is one option key.

Table 20 summarizes the benchmark prompt templates used for PTJ and VII. The prompt fields are instantiated from each sample's participant profile, transcript when available, scene context, previous-segment summary, and recommendation options. Task2.2 clips follow the shared benchmark video rendering described in Section E.5; they are generated from the original 1600×1200 recording as 640×480, 500k-bitrate, 1 FPS videos. Transcripts are generated for models without native audio input, using automatic speech recognition with Xunfei LFASR (iFLYTEK, 2026) for Chinese recordings and ElevenLabs Speech to Text (ElevenLabs) for English recordings.

Table 20: Prompt templates used for Task2.2 benchmark questions.

Question Type	Prompt Inputs	Output Constraint
PTJ	Fixation-overlaid clip; participant profile; transcript when available; interruptibility criteria covering safety-critical activity, focused work, deep social engagement, eating/drinking, appreciation or entertainment, transactions, privacy, rest, and device-boundary states.	Decide whether the user can be proactively interrupted during the segment. Output exactly one label: `True` if interruptible, or `False` if the segment should not be interrupted.
VII	Fixation-overlaid clip; participant profile; scene context; previous-segment summary; four recommendation options with intent categories.	Select the recommendation that should be proactively offered during the segment. Output exactly one option key from A/B/C/D, with no explanation.

J.2 Benchmark Statistic.

Table 21 summarizes the scale of the two Task2.2 benchmark question types. PTJ contains 269 binary samples from 269 unique fixation-overlaid clips across 15 participants. VII contains 1,161 four-choice samples from 742 unique fixation-overlaid clips across 15 participants, with one sample constructed for each participant-accepted proactive recommendation. Table 22 reports the annotation statistics before benchmark instance construction.

Table 21: Task2.2 benchmark statistics.

Question Type	#Samples	#Participants	#Unique Clips	Answer Space	Chance
PTJ	269	15	269	True / False	0.50
VII	1,161	15	742	A / B / C / D	0.25

PTJ has a moderately imbalanced binary label distribution (Table 23), with interruptible intervals forming the majority class. The benchmark preserves this imbalance to test whether proactive systems

Table 22: Original annotation statistics used to construct Task2.2 benchmarks.

Statistic	Scope	Value
PTJ original forbidden-segment annotations		
Total annotated forbidden segments	15 participants	74
Avg. forbidden segments / participant	15 participants	4.93
Avg. forbidden-segment length	74 segments	345.2 sec.
VII original merged recommendation annotations		
Total annotated segments	15 participants	519
Avg. segments / participant	15 participants	34.6
Avg. generated recommendations / segment	519 segments	25.9
Avg. accepted recommendations / segment	519 segments	2.24

can avoid disturbing users during participant-marked forbidden contexts. PTJ clips range from 19.0 to 300.0 seconds (mean 240.7 seconds; median 300.0 seconds). The transcript-bearing prompt version contains non-empty transcripts for 255/269 samples (94.8%) and is used for models without native audio understanding; native-audio models can instead consume the original clip audio directly.

Table 23: PTJ label distribution.

Label	#Samples	Percent
True (interruptible)	162	60.2%
False (do not interrupt)	107	39.8%

VII has a multi-class distribution over accepted recommendation types (Table 24) and a separately shuffled answer-key distribution (Table 25). Because several participant-accepted recommendations can occur within the same source segment, the 1,161 VII samples reuse 742 unique fixation-overlaid clips; among reused clips, each clip appears between 1 and 10 times (mean 1.56 samples per clip). Participant-annotated recommendation presentation time windows range from 0.8 to 300.0 seconds (mean 115.1 seconds; median 52.8 seconds). Trimmed clips range from 60.8 to 360.0 seconds (mean 175.1 seconds; median 112.8 seconds), reflecting the additional 60-second context prepended before each accepted recommendation. The transcript version contains non-empty transcripts for 1,129/1,161 samples (97.2%).

Table 24: VII accepted-recommendation type distribution. Each sample corresponds to one participant-accepted proactive recommendation.

Type	#Samples	Percent	Type	#Samples	Percent
Context-Aware Recommendation	252	21.7%	Computation and Estimation	108	9.3%
Task Assistance	185	15.9%	Decision Support	106	9.1%
Engaging Interaction	152	13.1%	Contextual Memory	88	7.6%
Procedural Guidance	123	10.6%	Translation, Text, and Audio Recognition	33	2.8%
Object Identification and Recognition	114	9.8%	–	–	–

J.3 Benchmark Results.

For reproducibility, Table 26 reports the main inference settings used for Task2.2. Vision-only and transcript-dependent video models are evaluated with the transcript-bearing prompt; native audio-video models are evaluated with the paired no-transcript prompt and consume the clip audio directly. The same decoding settings are used for both question types. Most non-thinking models use a maximum of 5 new tokens; qwen3.5-omni-plus requires at least 10 new tokens. Temperature is 0.0 for all models except InternVL3.5-8B-Instruct (thinking), which requires temperature 0.6 to enable thinking mode.

Table 25: VII correct option-key distribution.

Correct Key	#Samples	Percent	Correct Key	#Samples	Percent
A	280	24.1%	C	326	28.1%
B	266	22.9%	D	289	24.9%

Table 26: Task2.2 inference settings. "Frames/input" describes the visual/audio inputs supplied per clip; native-video API models handle frame selection internally. Gemini runs use full video and audio through the Files API without an explicit generation-token cap. Kimi and Qwen API runs start from the full video and dynamically reduce FPS/resolution to fit the API byte/token limit (20MB).

Models	Frames/input	Transcript or audio	Temp.	Max new tokens
Closed-source/API models				
gemini-3-flash-preview; gemini-3.1-pro-preview	native video + audio	audio	0.0	–
qwen3.5-omni-plus	native video + audio	audio	0.0	10
kimi-k2.5	native video	transcript	0.0	5
kimi-k2.5 (thinking); qwen3.5-397b-a17b	native video	transcript	0.0	8192
gpt-4o	50 frames	transcript	0.0	5
Open-source/local models				
Qwen3-Omni-30B-A3B-Instruct	native video + audio	audio	0.0	5
Qwen3-Omni-30B-A3B-Thinking	native video + audio	audio	0.0	8192
MiniCPM-V-4_5	native video ($\leq$180 frames)	transcript	0.0	5
Qwen2.5-VL-7B-Instruct; Qwen3-VL-8B-Instruct	native video	transcript	0.0	5
Qwen3-VL-8B-Thinking	native video	transcript	0.0	8192
LongVA-7B-DPO	360 frames	transcript	0.0	5
InternVL3-8B; InternVL3_5-8B-Instruct	135 frames	transcript	0.0	5
InternVL3_5-8B-Instruct (thinking)	135 frames	transcript	0.6	8192
llava-onevision-qwen2-7b-ov-hf; LongVU_Qwen2_7B	64 frames	transcript	0.0	5
InternVL2_5-8B; LLaVA-Video-7B-Qwen2	45 frames	transcript	0.0	5

Table 27: Task 2.2 benchmark results under the transcript/native-audio setting. PTJ: random chance 50.00%. VII: random chance 25.00%. Scores are percentages ($\times 100$). **Bold** marks the best score in each numeric column within the Closed-source/API and Open-source/local blocks.

Models	PTJ				VII	
	acc	**macro f1**	f1 (true)	f1 (false)	**acc**	macro f1
Closed-source/API models						
gemini-3.1-pro-preview	57.25	57.13	54.90	59.36	**55.81**	**55.76**
gemini-3-flash-preview	57.62	57.41	54.40	60.42	52.80	52.66
kimi-k2.5	**59.85**	**59.78**	**61.43**	58.14	54.78	54.66
kimi-k2.5 (thinking)	54.28	54.21	52.51	55.91	52.37	52.08
qwen3.5-397b-a17b	58.36	58.36	58.52	58.21	51.42	51.28
qwen3.5-omni-plus	54.65	53.22	45.05	**61.39**	49.96	49.44
gpt-4o[†]	37.92	29.55	1.22	57.88	43.84	44.94
Open-source/local models						
InternVL2_5-8B	66.54	59.06	76.56	41.56	43.58	43.27
InternVL3-8B	60.59	56.22	70.06	42.39	42.89	42.34
InternVL3_5-8B-Instruct	53.53	52.78	46.81	**58.75**	41.69	41.02
InternVL3_5-8B-Instruct (thinking)	46.84	46.70	43.92	49.47	41.26	41.03
MiniCPM-V-4_5	60.22	60.00	62.98	57.03	43.32	43.13
Qwen2.5-VL-7B-Instruct	39.41	28.27	0.00	56.53	41.09	39.33
Qwen3-Omni-30B-A3B-Instruct	**66.91**	57.62	**77.47**	37.76	46.34	45.36
Qwen3-Omni-30B-A3B-Thinking	60.97	59.92	66.88	52.97	47.29	46.62
Qwen3-VL-8B-Instruct	65.80	**60.69**	74.86	46.51	**47.63**	**47.16**
Qwen3-VL-8B-Thinking	61.71	57.57	70.82	44.32	46.17	45.33
LongVU_Qwen2_7B	60.22	37.59	75.17	0.00	37.98	35.97
LLaVA-Video-7B-Qwen2	60.97	39.60	75.52	3.67	40.65	37.69
llava-onevision-qwen2-7b-ov-hf	57.62	56.97	62.25	51.69	38.50	36.72
LongVA-7B-DPO	60.22	38.45	75.06	1.83	36.09	34.86
human performance	68.97	68.36	72.73	64.00	68.00	70.34

[†] gpt-4o produced 25 invalid PTJ predictions (9.3%) and 89 invalid VII predictions (7.7%) due to content-safety refusals.

J.3.1 BASELINE RESULTS

PTJ is evaluated primarily with macro F1 because the binary labels are moderately imbalanced and a useful proactive assistant must recognize both interruptible and non-interruptible contexts. Among closed-source/API models, kimi-k2.5 obtains the best macro F1 (59.78), while qwen3.5-omni-plus achieves the highest F1 on the non-interruptible class but at the cost of lower F1 on interruptible clips. Among open-source/local models, Qwen3-VL-8B-Instruct reaches the best macro F1 (60.69), whereas Qwen3-Omni-30B-A3B-Instruct obtains the highest accuracy and F1 on interruptible clips. The per-class scores reveal different failure modes: LongVU, LLaVA-Video, and LongVA strongly favor the majority `True` label and miss many participant-marked forbidden contexts, while Qwen2.5-VL-7B-Instruct collapses toward `False`, producing zero F1 on interruptible clips. MiniCPM-V-4.5 is notable for a more balanced PTJ profile despite lower absolute accuracy.

VII is evaluated primarily with accuracy because the shuffled multiple-choice labels are approximately balanced. This semantic selection task requires the model to distinguish the participant-accepted proactive request from plausible alternatives generated for the same segment. Gemini-3.1-pro-preview performs best overall (55.81 accuracy), followed by kimi-k2.5 and gemini-3-flash-preview among closed-source/API models. The best open-source/local result is Qwen3-VL-8B-Instruct (47.63 accuracy), leaving a substantial gap to the strongest API models. Native audio does not uniformly improve results. Audio-capable omni models are competitive but do not consistently outperform transcript-conditioned MLLMs, suggesting that explicit transcripts can capture much of the audio information needed for this task. Thinking modes are also not consistently beneficial under the constrained answer format, likely because both PTJ and VII require short-form decisions rather than long free-form reasoning.

We additionally conducted a human prediction study to test whether the participant-provided Task2.2 labels are predictable from the same inputs given to models and can be approximated through human annotation. For each question type, we sampled a label-stratified and participant-balanced subset from the full test split, targeting around 10% of the data with a maximum of 50 examples per task. This yielded 29 PTJ examples and 50 VII examples, evaluated using the same inputs and answer formats as the model benchmark. Because the ground truth reflects the original participant's retrospective interruption judgments and recommendation preferences, the study should not be read as an absolute human upper bound. Instead, it measures cross-person predictability from the available video, dialogue, participant profile, and candidate recommendations. Human annotators substantially outperform the strongest models on the primary metrics, reaching 68.36 macro F1 on PTJ and 68.00 accuracy on VII. These results support that Task2.2 labels are not arbitrary despite their subjective origin. With careful inspection of the scene, dialogue, and participant profile, humans can often infer when interruption is acceptable and which proactive request the participant would prefer. The remaining gap between human and perfect performance still reflects genuine personalization difficulty, but the human advantage over MLLMs indicates that the benchmark captures predictable user-aligned judgments rather than unknowable preferences.

J.3.2 ERROR ANALYSIS

We analyze errors for kimi-k2.5 and Qwen3-VL-8B-Instruct, which are the strongest representative closed-source/API and open-source/local models under the transcript setting.

PTJ errors reflect different interruption biases. The two models reach similar PTJ macro F1, but make qualitatively different mistakes, as shown in Figure 16. Kimi-k2.5 predicts `False` more often than `True`, with 151/269 predictions (56.1%) versus 118/269 predictions (43.9%), correctly identifying 75 of 107 non-interruptible segments (70.1%) but missing 76 of 162 interruptible segments (46.9%). This pattern suggests a conservative bias that avoids interruption in many ambiguous cases. Qwen3-VL-8B-Instruct shows the opposite tendency: it predicts `True` in 204 of 269 cases (75.8%), correctly recovering 137 of 162 interruptible segments (84.6%) but incorrectly interrupting on 67 of 107 non-interruptible segments (62.6%). Thus, the best open-source/local model obtains higher PTJ accuracy by capturing the majority interruptible label, while kimi-k2.5 is more balanced in recognizing participant-marked forbidden contexts. The models are jointly wrong on only 49 PTJ segments (18.2%), indicating that their errors are partly complementary rather than identical.

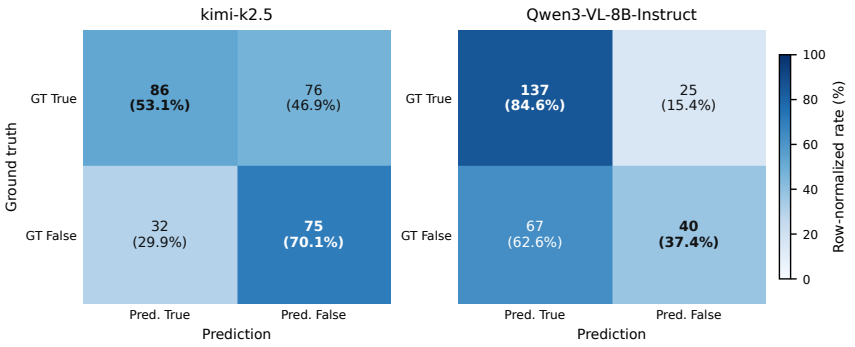


Figure 16: PTJ confusion heatmaps for the two representative models. Each cell reports count and row-normalized percentage. The `True` label denotes an interruptible segment, and `False` denotes a participant-marked non-interruptible segment.

VII errors expose hard preference distinctions. For VII, kimi-k2.5 is more accurate than Qwen3-VL-8B-Instruct (636 vs. 553 correct predictions out of 1,161), but the two models share many difficult cases. Figure 17 summarizes their prediction overlap and type-specific accuracy. They are both correct on 439 instances and both wrong on 411 instances; among the shared errors, they choose the same wrong option in 296 cases (72.0%). This suggests that many VII failures arise from highly plausible distractors generated for the same segment, not only from random answer-key bias. This suggests that many VII failures arise from highly plausible distractors generated for the same segment. This pattern is expected from the construction of VII, since distractors are rejected recommendations generated for the same segment, so they are often visually and contextually plausible but fail to match the participant's preference. The shuffled ground-truth positions are approximately balanced, but Qwen3-VL-8B-Instruct over-selects option D (394/1,161, 33.9%), whereas kimi-k2.5 remains closer to the balanced key distribution. Errors also vary by recommendation type. Both models perform best on context-aware recommendation and engaging interaction examples, with 73.8% and 63.8% accuracy for kimi-k2.5 and 61.9% and 60.5% for Qwen3-VL-8B-Instruct. In contrast, computation/estimation and contextual memory are among the hardest categories, with 25.9% and 26.1% accuracy for kimi-k2.5 and 22.2% and 30.7% for Qwen3-VL-8B-Instruct. These harder cases require the model to infer a participant's specific informational need or remembered context, which is often less visually explicit than recognizing an immediately relevant scene-based suggestion.

Human-correct cases reveal missing contextual judgment. We further inspect cases in the transcript-based human prediction subset where human annotators answer correctly but both representative models fail. This occurs in 4 of 29 PTJ examples and 8 of 50 VII examples. Figure 18 shows one example from each question type. In the PTJ example, both models classify the segment as non-interruptible, likely because the user is engaged in a hands-on activity. However, the broader context suggests low-risk self-practice guided by video, making interruption acceptable. In the VII example, both models choose a plausible screen-level decision about game mode, but the participant-accepted answer asks which key the shopkeeper said was for jumping. Human annotators can reason from the user's dialogue and movement that the participant is still unfamiliar with the console, controls, and game, making it more useful to recall the clerk's recent instruction than to recommend a game-mode choice. The failure reflects preference alignment and contextual memory rather than scene understanding alone: the rejected option is relevant to the visible game screen, but the prior dialogue makes the remembered control instruction the more useful intervention.

J.3.3 Ablation Study

Table 28 isolates the contribution of the main input sources used in Task2.2. We compare plain video, fixation-overlaid video, fixation-overlaid video with transcript, and fixation-overlaid video with transcript and in-context learning (ICL). The ICL condition serves as a lightweight personalization signal rather than a generic few-shot prompt: for VII, each participant is represented by one randomly selected example of their accepted recommendation behavior; for PTJ, each participant contributes one interruptible and one non-interruptible example when both are available. These examples are

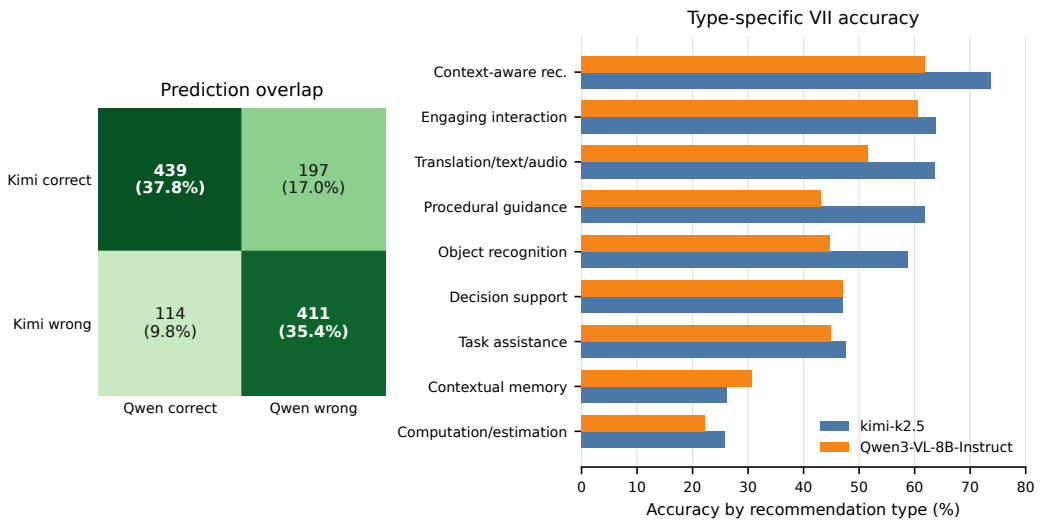


Figure 17: VII error diagnostics for kimi-k2.5 and Qwen3-VL-8B-Instruct. The left heatmap shows whether the two models succeed or fail on the same multiple-choice instances. The right panel compares accuracy across recommendation types, revealing which preference distinctions remain difficult even for the strongest representative models.

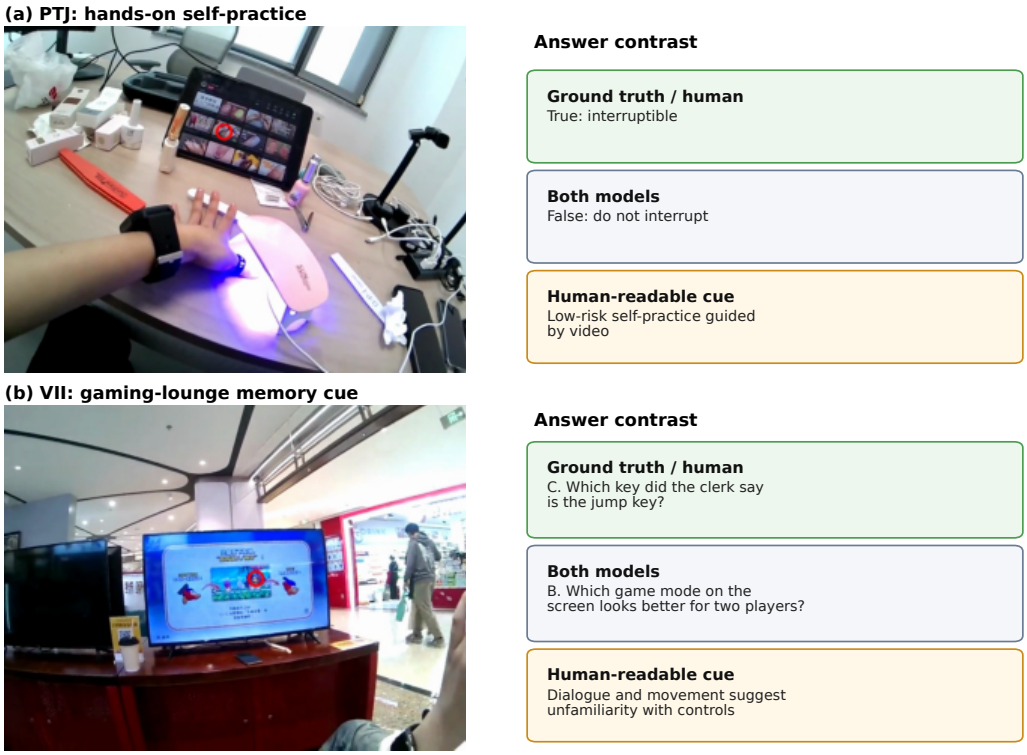


Figure 18: Representative human-correct/model-wrong Task2.2 examples. Each panel shows a rendered benchmark frame and the answer contrast between human annotators, ground truth, and the two representative models. The examples illustrate that humans can use broader activity context and participant intent to resolve cases where both models select conservative or plausible but preference-misaligned answers.

removed from the evaluated ICL split, so the comparison measures whether user-specific history improves prediction on the remaining samples.

Table 28: Task 2.2 ablation over input conditioning for representative closed-source/API and open-source/local models (same decoding as Table 27). Scores are percentages ($\times 100$). ICL denotes same-participant in-context examples used as past interaction history. **Bold** marks the best score in each numeric column within each model block.

Model	Inputs	PTJ				VII	
		acc	**macro f1**	f1 (true)	f1 (false)	**acc**	macro f1
Closed-source/API models							
kimi-k2.5	video	60.22	59.90	63.48	56.33	47.29	47.15
	video + gaze	57.99	57.96	59.21	56.70	46.25	46.20
	video + gaze + transcript	59.85	59.78	61.43	58.14	54.78	54.66
	video + gaze + transcript + ICL	**68.33**	**67.83**	**71.85**	**63.83**	†**54.89**	†**54.85**
Open-source/local models							
Qwen3-VL-8B-Instruct	video	66.54	59.37	76.44	42.31	42.72	41.80
	video + gaze	61.71	56.38	71.63	41.14	42.29	41.42
	video + gaze + transcript	65.80	60.69	74.86	**46.51**	47.63	47.16
	video + gaze + transcript + ICL	**68.33**	**61.35**	**77.78**	44.93	**48.95**	**48.10**

†The VII score for kimi-k2.5 uses a multi-turn dialogue format for the ICL example, which performed better for this model than the standard single-turn prompt format used by the remaining ICL runs.

Transcripts provide the clearest contribution. Task2.2 benefits most consistently from linguistic context. Adding transcripts to fixation-overlaid video produces large gains for VII, increasing accuracy by **8.5 points** for kimi-k2.5 (46.25 to 54.78) and **5.3 points** for Qwen3-VL-8B-Instruct (42.29 to 47.63). It also improves PTJ macro F1 relative to the gaze-only setting, especially for Qwen3-VL-8B-Instruct (56.38 to 60.69). These gains indicate that proactive request prediction depends strongly on audio-derived information, including the participant's speech and surrounding sounds, especially for models without native audio understanding.

Same-participant ICL improves personalization. Adding ICL further improves both representative models. The largest effect appears on PTJ for kimi-k2.5, where macro F1 rises by **8.1 points** (59.78 to 67.83). Qwen3-VL-8B-Instruct shows smaller but consistent gains, improving from 60.69 to 61.35 macro F1 on PTJ and from 47.63 to 48.95 accuracy on VII. Because the examples come from the same participant, these gains support the value of user-specific history: past judgments help the model estimate interruption tolerance and preferred proactive assistance beyond what can be inferred from the current segment alone.

Fixation overlays alone do not reliably help current MLLMs. Although fixation cues should intuitively help models reason about the user's visual attention, the gaze-only comparison is more nuanced. For both representative models, fixation-overlaid video slightly underperforms plain video before transcripts are added: kimi-k2.5 drops from 59.90 to 57.96 macro F1 on PTJ and from 47.29 to 46.25 accuracy on VII, while Qwen3-VL-8B-Instruct drops from 59.37 to 56.38 macro F1 on PTJ and from 42.72 to 42.29 accuracy on VII. Prior work shows that red-circle markings can steer VLM attention in static grounding tasks (Shtedritski et al., 2023), and EgoGazeVQA demonstrates that gaze-guided prompting can improve MLLM performance on egocentric video questions requiring precise spatial reasoning and intent interpretation (Peng et al., 2025). However, EgoGazeVQA also reports that gaze cues can fail under heavy camera/body motion and that rapid gaze transitions may direct models toward irrelevant objects. These observations align with our setting, where the fixation marker is a temporally varying attention signal sampled in a few-minute egocentric clip rather than a single object-level visual prompt, and the target labels are participant-specific judgments of interruptibility or recommendation value rather than direct object localization. At the 1 FPS benchmark sampling rate, fixation cues may be too sparse or visually distracting for models to aggregate reliably, and many ground-truth decisions may depend more on speech, activity context, or personal preference than on the exact fixation point. Thus, the ablation suggests that fixation overlays alone are not sufficient for current general-purpose MLLMs to exploit gaze robustly in Task2.2. Future work may require higher-rate temporal gaze modeling, gaze-aware fine-tuning, or architectures explicitly trained to interpret fixation overlays as user-attention signals.

K BENCHMARK TASK3.1 - MEMORY RECALL PREDICTION

K.1 BENCHMARK QUESTION DEFINITION AND CURATION PIPELINE.

As introduced in Section 4.4.1, the MRP task evaluates a model's capacity to simulate human memory patterns by assessing its ability to identify details that are most likely to be clearly recalled by a human observer.

The MRP is formulated as a four-choice multiple-choice question. Each sample comprises a fixation-overlaid video clip with audio track, scene context, a transcript (or audio), and a set of four options: $\mathcal{C}_i = \{c_A, c_B, c_C, c_D\}$. The model must predict a single option key $\hat{y}_i \in \{A, B, C, D\}$. The correct option is derived from a participant-annotated vivid detail, while the distractors consist of other items or events that appear within the same video segment

The construction of MRP follows a rigorous multi-stage pipeline:

- **Textual Refinement:** We begin with the original participant accounts. Accounts that are overly dense, excessively colloquial, or too lengthy are manually refined into a formal, concise format suitable for standardized options.
- **Segment Sampling and Generation:** For each vivid detail, we programmatically extract a 3-minute video clip centered around the relevant timestamp. We then utilize Gemini to generate three distractors based on other elements present in the clip. To ensure linguistic consistency, the model is prompted to match the style of the original vivid detail. Furthermore, the description of the vivid detail is subtly adjusted to align with the phrasing and complexity of the generated distractors.
- **Quality Control:** Experimenters manually curate the candidate set to maintain an appropriate difficulty balance. We eliminate distractors that are either too trivial (e.g., common road signs) or excessively obscure (e.g., objects that are not visually salient). We also conduct a final verification to ensure that the semantic integrity of the correct option is preserved after LLM-based adjustment.
- **Collision Avoidance:** In cases where vivid details are temporally close and 3-minute clips inevitably overlap, we explicitly instruct the API to exclude other annotated vivid details from the distractor set. These instances are further cross-checked manually by experimenters.
- **Temporal Shuffling**:To ensure that models rely on true multimodal understanding rather than temporal heuristics, the candidate options are randomly shuffled. This ensures that the order of options does not correspond to the chronological sequence of events in the video, thereby requiring the model to engage deeply with the visual and narrative context.

Table 29 summarizes the benchmark prompt templates used for MRP. The prompt field are instantiated from each sample's scene contex,transcripts when available and option choices. transcripts are only provided for models cannot deal with audio.

Table 29: Prompt templates used for MRP benchmark questions.

Question Type	Prompt Inputs	Output Constraint
MRP	Fixation-overlaid clip; scene context; transcript when available, four detail options .	Select the detail that is most recallable and most likely to stay vivid in memory after the event. Output exactly one option key from `A`/`B`/`C`/`D`, with no explanation.

K.2 BENCHMARK STATISTIC.

The MRP dataset consists of 200 samples derived from 200 distinct video clips across 15 participants, where each sample corresponds to a single annotated vivid detail. The distribution of annotations per participant ranges from 3 to 23, with a mean (μ) of 13.33 and a median of 12. To maintain consistency in temporal context, all video clips are standardized to a duration of 180 seconds.

K.3 BENCHMARK RESULTS.

For reproducibility, Table 30 reports the main inference settings used for Task3.1. Vision-only and transcript-dependent video models are evaluated with the transcript-bearing benchmark files; native audio-video models are evaluated with the paired no-transcript files and consume the clip audio directly. The same decoding settings are used for both question types. Most non-thinking models use a maximum of 5 new tokens; qwen3.5-omni-plus requires at least 10 new tokens. Temperature is 0.0 for all models except InternVL3.5-8B-Instruct (thinking), which requires temperature 0.6 to enable thinking mode.

Table 30: Task3.1 inference settings. "Frames/input" describes the visual/audio inputs supplied per clip; native-video API models handle frame selection internally. Gemini runs use full video and audio through the Files API without an explicit generation-token cap. Kimi and Qwen API runs start from the full video and dynamically reduce FPS/resolution to fit the API byte/token limit (20MB).

Models	Frames/input	Transcript or audio	Temp.	Max new tokens
Closed-source/API models				
gemini-3-flash-preview; gemini-3.1-pro-preview	native video + audio	audio	0.0	–
qwen3.5-omni-plus	native video + audio	audio	0.0	10
kimi-k2.5	native video	transcript	0.0	5
kimi-k2.5 (thinking); qwen3.5-397b-a17b	native video	transcript	0.0	8192
gpt-4o	50 frames	transcript	0.0	5
Open-source/local models				
Qwen3-Omni-30B-A3B-Instruct	native video + audio	audio	0.0	5
Qwen3-Omni-30B-A3B-Thinking	native video + audio	audio	0.6	8192
MiniCPM-V-4_5	native video ($\leq$180 frames)	transcript	0.0	32
Qwen2.5-VL-7B-Instruct	native video	transcript	0.0	5
Qwen3-VL-8B-Instruct	native video	transcript	0.0	32
Qwen3-VL-8B-Thinking	native video	transcript	0.0	2048
LongVA-7B-DPO	360 frames	transcript	0.0	5
InternVL3-8B; InternVL3_5-8B-Instruct	150 frames	transcript	0.0	5
InternVL3_5-8B-Instruct (thinking)	150 frames	transcript	0.6	2048
llava-onevision-qwen2-7b-ov-hf; LongVU_Qwen2_7B	64 frames	transcript	0.0	5
InternVL2_5-8B	58 frames	transcript	0.0	5
LLaVA-Video-7B-Qwen2	64 frames	transcript	0.0	5

K.3.1 BASELINE RESULTS

As a semantic selection task, Memory Recall Prediction (MRP) requires models to distinguish vivid details recalled by participants from plausible distractors derived from the same segments. Detailed in Table 31, our experimental results show that among closed-source and API-based models, *gpt-4o* achieves the best overall performance with an accuracy of 51.50% and a macro F1 of 44.01, closely followed by *qwen3.5-omni-plus* (52.00 accuracy, 42.85 macro F1). In the open-source and local category, *Qwen3-Omni-30B-A3B-Instruct* delivers the highest accuracy at 55.50%, though its macro F1 (38.27) indicates a more skewed prediction profile. Notably, reasoning-heavy "thinking" models frequently underperform compared to their instruction-tuned counterparts; for instance, *Qwen3-Omni-30B-A3B-Thinking* shows a significant drop in accuracy (47.50%) relative to its base instruct version. This performance discrepancy likely stems from the intuitive nature of vivid human memory; excessive deliberation within these reasoning-heavy architectures may inadvertently obscure the subtle, salient nuances prioritized by human recall, leading to suboptimal selection.

Regarding the **human expert baseline**, as indicated in Table 31, there remains a substantial performance gap between the most advanced MLLMs and human experts across almost all cognitive dimensions. Human experts achieve an accuracy of 76.67% in the MRP task, significantly outperforming the best-performing model, *Qwen3-Omni-30B-A3B-Instruct*, by over 21%.

K.3.2 ABLATION STUDY

Table 32 presents the ablation study results for the *Memory Recall Prediction* (MRP) task, evaluating the performance of both closed-source (*Kimi-k2.5*) and open-source (*Qwen3-VL-8B-Instruct*) models under various input configurations. The analysis provides several critical insights into how multimodal cues and prompting strategies contribute to modeling memory reachability:

Table 31: Benchmark Results for Cognitive Memory

Models	MRP		MIM-item-text		MIM-item-photo		MIM-Lifespan			
	acc	macro f1	acc	macro f1	acc	macro f1	acc	macro f1	f1($\leq 24h$)	f1($> 24h$)
Closed-source models										
Gemini-3.1-pro-preview	50.00	41.06	47.06	43.03	**64.10**	**63.21**	**85.39**	**82.34**	**75.00**	**89.68**
Gemini-3-flash-preview	47.00	38.71	**53.85**	**51.39**	43.14	38.55	79.76	74.33	62.50	86.15
gpt-4o	51.50	**44.01**	47.06	43.55	57.89	56.75	73.03	71.75	63.49	80.0
kimi-k2.5	47.50	37.03	49.02	45.53	51.28	51.65	78.09	74.89	62.96	84.04
kimi-k2.5(thinking)	48.50	38.61	45.10	40.15	48.72	48.33	82.02	76.55	65.22	88.21
qwen3.5-397b-a17b	49.00	40.33	39.22	34.51	51.28	49.87	79.21	74.60	63.36	85.83
qwen3.5-omni-plus	**52.00**	42.85	45.10	38.32	61.54	57.65	82.02	78.60	69.90	87.30
Open-source models										
InternVL2_5-8B	43.00	35.95	40.00	36.46	46.15	**41.38**	53.37	52.71	47.13	58.29
InternVL3-8B	40.50	34.05	29.41	28.86	43.59	38.70	**75.84**	**71.86**	**61.26**	**84.44**
InternVL3_5-8B-Instruct	35.50	28.74	43.14	38.97	41.03	38.67	48.31	48.29	49.45	47.13
InternVL3_5-8B-Thinking	34.50	28.47	**56.86**	**52.83**	35.90	35.39	50.56	55.55	50.00	51.11
MiniCPM-V-4_5	50.00	39.44	31.37	29.40	15.38	10.26	69.66	67.06	57.81	76.32
Qwen2.5-VL-7B-Instruct	38.00	29.91	33.33	28.93	33.33	33.11	48.31	48.29	47.14	49.45
Qwen3-Omni-30B-A3B-Instruct	**55.50**	38.27	29.41	25.36	35.90	36.30	61.80	60.52	53.42	67.62
Qwen3-Omni-30B-A3B-Thinking	47.50	35.99	29.41	26.78	41.03	38.19	75.28	71.06	60.00	82.11
Qwen3-VL-8B-Instruct	47.50	39.49	47.06	41.99	38.46	31.63	47.78	47.75	46.75	48.75
Qwen3-VL-8B-Thinking	43.00	34.23	37.25	36.11	31.58	31.15	52.25	52.23	51.43	53.04
LongVU_Qwen2_7B	36.00	31.75	27.45	22.64	**48.72**	24.83	13.48	22.45	25.97	18.92
LongVA-7B-DPO	49.00	**41.43**	37.25	32.16	25.64	15.48	24.16	19.13	39.27	0.00
llava-onevision-qwen2-7b-ov-hf	47.00	40.25	29.41	27.14	21.62	14.38	25.84	20.54	41.07	0.00
LLaVA-Video-7B-Qwen2	42.50	35.37	25.49	20.87	23.08	9.57	42.14	42.12	43.09	41.14
human performance	76.67	40.11	60.00	61.11	75.00	66.67	88.89	87.50	83.33	91.67

Table 32: Ablation results for MRP. Bold indicates the best performance within each model category.

Model	Inputs	MRP	
		acc	macro f1
Closed-source/API models			
kimi-k2.5	video	45.00	37.81
	video + gaze	43.00	35.33
	video + gaze + transcript	47.50	37.03
	video + gaze + transcript + ICL	**56.22**	**45.83**
Open-source/local models			
Qwen3-VL-8B-Instruct	video	37.50	31.48
	video + gaze	35.00	29.16
	video + gaze + transcript	47.50	**39.49**
	video + gaze + transcript + ICL	**48.65**	37.54

- **The Pivotal Role of Transcripts:** The integration of *transcripts*—textual transcriptions of the video's audio track—emerges as a primary driver of performance gains. For *Qwen3-VL-8B-Instruct*, adding transcripts to the video and gaze baseline yields a significant improvement, increasing Accuracy from 35.00% to 47.50% and Macro F1 from 29.16 to 39.49. This suggests that linguistic information provides essential semantic context that helps the model ground vivid details within the associated past events.
- **Synergy between Transcripts and ICL:** The highest overall performance is consistently achieved by integrating all modalities with *In-Context Learning* (ICL). Specifically, *Kimi-k2.5* reaches its peak performance (56.22% Accuracy and 45.83 Macro F1) under the full configuration. These results suggest that while transcripts provide the essential factual grounding, ICL serves as a powerful calibration mechanism. By presenting one-shot example of a user's recall, ICL enables the model to characterize the individual's unique **memory patterns**—specifically how they prioritize certain vivid details while others succumb to **memory decay**. This allows the model to better navigate the complex mapping between multimodal cues and the subjective nature of human memory retrieval.
- **Observations on Gaze Information:** Interestingly, the isolated addition of gaze signals g_i to the video-only baseline leads to a marginal performance dip for both models. For instance, *Kimi-k2.5*'s accuracy decreases from 45.00% to 43.00% when gaze is introduced without textual context. This suggests that raw gaze duration may be a misleading indicator

of memory salience; for example, a prolonged fixation on complex text may reflect high cognitive load rather than successful encoding, whereas a transient glance at a distinct object might result in a lasting memory.

- **Model Comparison:** Across all configurations, the closed-source *Kimi-k2.5* generally maintains a performance lead over the open-source *Qwen3-VL-8B-Instruct*, particularly in its ability to leverage one-shot examples through ICL to resolve the most challenging memory recall queries.

In summary, the results underscore that transcripts are indispensable for bridging the gap between raw egocentric video v_i and high-level cognitive recall, while ICL serves as a powerful mechanism to refine the model's predictive accuracy in the MRP task.

L Benchmark Task 3.2 - Memory Intent Modeling

L.1 Benchmark Question Definition and Curation Pipeline.

As described in section 4.4.2, MIM evaluates the model's understanding of memory intent from two aspects: whether it can tell the item a user explicitly intends to preserve for future utility and whether it can understand the temporal relevance of information. We instantiate these perspectives as two benchmark question types.

Memory Assistance Recognition(MAR). MAR is a four choice multi-choice question task that asks the model select the item the user want to help remember. Each sample comprises a fixation-overlaid video clip with audio track, scene context, a transcript (or audio), and a set of four options: $\mathcal{C}_i = \{c_A, c_B, c_C, c_D\}$. The model must predict a single option key $\hat{y}_i \in \{A, B, C, D\}$. The correct option is derived from a participant-annotated memory content, while the distractors consist of other items or events that appear within the same video segment.

The construction of MAR follows a rigorous multi-stage pipeline:

- **Type-Aware Segment Construction:** Because memory intent is instantiated as three sub-types—short-term memory (STM), long-term event (LTE), and long-term object (LTO)—we extract a clip whose temporal extent reflects the granularity of each type: STM uses the annotated cue interval (extended to a minimum of 12 s), LTE uses the full event interval, and LTO uses a ± 15 s window around the annotation timestamp. Any segment exceeding 45 s is re-centered on the interval midpoint and truncated to a 45 s window, keeping the cue moment at the center of the clip. MAR does not retain audio tracks that explicitly reveal the answer or whose memory target is itself audio; all other audio cues are removed to prevent shortcut solutions while keeping the task focused on visual and contextual reasoning.
- **Textual Refinement:** Raw participant accounts are frequently elliptical and rely on private context. We apply a type-specific LLM rewriting pass that resolves deictic references and normalizes phrasing: STM/LTE rewrites preserve the "item-to-remember" framing while expanding vague references using the annotator-provided memory reason, whereas LTO rewrites collapse the account to the target object itself, stripping the motivation for remembering it.
- **Segment Sampling and Generation:** Distractors are grounded in the same recording session as the target. For each question we sample three non-overlapping sub-clips from a local window around the target timestamp, enforcing a minimum temporal separation to guarantee visual diversity. Each sub-clip is then converted into one option: for text-form questions, Gemini is prompted to generate a textual description of the sub-clip conditioned on scene context (location, participants, main activities, emotion), so that the distractor text mimics the density and framing of the correct option; for image-form questions, a reference frame is extracted from the sub-clip and used directly as the option visual.
- **Quality Control and Collision Avoidance:** Experimenters manually curate the candidate set to remove options that are trivially identifiable or not visually salient, and verify that LLM-based rewriting preserves the semantic identity of the correct option. For image-form questions, experimenters additionally draw a bounding box around the candidate object on each reference frame, so that the visual referent of every option is unambiguous. When

multiple annotated memory items fall within the same local window, their intervals are explicitly excluded from the distractor candidate region, and a greedy non-overlap placement procedure iteratively shrinks the sampling windows so that no distractor source overlaps another annotated memory item; these cases are further cross-checked manually.

- **Temporal Shuffling:** The four options are randomly shuffled before export, so that the position of the correct answer carries no information about its source and models cannot rely on positional or chronological heuristics.

Table 33 summarizes the benchmark prompt templates used for MAR. The prompt fields are instantiated from each sample's scene context, transcripts when available, and option choices.

Table 33: Prompt templates used for MAR benchmark questions.

Question Type	Prompt Inputs	Output Constraint
MAR	Fixation-overlaid clip with audio; scene context; four memory-item options.	Select the item the user most likely intends to preserve for future utility. Output exactly one option key from `A`/`B`/`C`/`D`, with no explanation.

Memory Lifespan Identification (MLI). MLI is a binary classification task designed to evaluate a model's ability to determine the temporal relevance of a memory item. Each sample consists of a fixation-overlaid video clip (without audio) and a classification guideline. Models are required to employ Chain-of-Thought (CoT) prompting, first outputting their reasoning before concluding with a definitive label: *within_24h* or *more_than_24h*.

The video segments for MLI are derived from user-annotated memory content through a rigorous curation process:

- **Standardization and Truncation:** For annotated segments exceeding 15 seconds, experimenters manually verify the content. If the core memory information remains consistent throughout, the segment is truncated into a 15-second clip. Any remaining fragments shorter than 15 seconds are discarded.
- **Information Integrity:** A primary constraint during curation is the preservation of "memory needs." If truncating a segment to 15 seconds results in severe information loss or renders the memory content unrecognizable, the original length is maintained to ensure sufficient context.
- **Context Extension:** For segments originally shorter than 15 seconds, experimenters manually review and extend the temporal boundaries to include necessary environmental context, ensuring the visual data is interpretable for the model.
- **Audio Redaction:** Because users performed *in-situ labeling* for the MIM dataset, the audio tracks often contain verbal cues or reasoning that explicitly reveal the information's lifespan. To prevent data leakage and ensure the model relies solely on visual and contextual cues, all audio tracks are removed.

The labels for this task are derived directly from the users' original temporal annotations.

Table 34 summarizes the benchmark prompt templates used for the Memory Lifespan Identification (MLI) task. The prompt configurations include specific classification examples designed to calibrate the model's understanding of user intent across various scenarios, such as utility tracking and knowledge acquisition. To guide the model through the complex cognitive process of judging memory utility, we implement a Chain-of-Thought (CoT) analysis protocol defined as follows:

- **Step 1: [WHAT]** Identify the specific content or intent the user is capturing in the egocentric video.
- **Step 2: [CONTEXT]** Determine the underlying motivation for capturing this information based on the established Scenario Guidelines.
- **Step 3: [REUSABILITY]** Assess whether the visual information retains long-term value for the user's future self or personal narrative once the immediate task concludes.

Table 34: Prompt templates and output constraints used for the MLI benchmark questions.

Task	Prompt Inputs	Output Constraint
MLI	Fixation-overlaid video clip; classification examples covering utility tracking, step-by-step execution, temporary status, journeys, knowledge acquisition, and sensory appreciation.	Provide reasoning following the three-step CoT guideline; output exactly `within_24h` or `more_than_24h` on the final line.

L.2 Benchmark Statistic.

Memory Assistance Recognition (MAR). The MAR benchmark comprises 90 samples collected from 15 participants. Following the type-aware curation pipeline, each sample is presented in one of two question forms: *text* questions, where the four options are pure text strings, and *image* questions, where each option is additionally anchored by an extracted reference frame with a human-drawn bounding box. Of the 90 samples, 51 are *text* questions and 39 are *image* questions (detailed in Table 35). To provide sufficient context for identifying the user's memory intent, the fixation-overlaid clips range from 119.4 s to 337.0 s, centered around a mean of 234.1 s and a median of 237.2 s. Detailed dataset specifications, including participant-wise annotation density and temporal metrics, are summarized in Table 36.

Table 35: MAR question-form distribution.

Question Form	#Samples	Percent
text	51	56.7%
image	39	43.3%

Table 36: Detailed statistics for MAR video clips and annotations.

Metric	Mean (μ)	Median	Min	Max
Annotations per Participant	6.0	4.0	1	17
Video Clip Duration (s)	234.1	237.2	119.4	337.0

For reproducibility, Table 40 reports the main inference settings used for MAR. Prompts require the model to select a single option key from {A,B,C,D} with no explanation. API models use native video where supported; local vLLM video models sample the clip per each runner's `video_fps` / frame recipe.

Memory Lifespan Identification(MLI). The MLI benchmark comprises 178 samples collected from 15 participants. The dataset exhibits a natural imbalance reflecting real-world memory patterns: 132 samples are labeled as *more_than_24h*, while 46 samples are labeled as *within_24h* (detailed in Table 38).To ensure sufficient context for temporal reasoning, the video clips vary in duration, ranging from 5.6s to 31.4s. The distribution is centered around a mean of 14.79s and a median of 15.0s, aligning with our 15-second curation pipeline. Detailed dataset specifications, including participant-wise annotation density and temporal metrics, are summarized in Table 39.

For reproducibility, Table 40 reports the main inference settings used for LFI. Prompts require chain-of-thought before the final label. API models use native video where supported; local vLLM video models sample the clip per each runner's `video_fps` / frame recipe.

L.3 Benchmark Results.

L.3.1 Baseline Results

MAR is evaluated with accuracy and macro F1 across two question forms—MIM-item-text (four purely textual options, primarily from short-term cues) and MIM-item-photo (four bbox-anchored reference-frame options, primarily for long-term object recognition)—because the two forms probe

Table 37: MAR inference settings. "Frames/input" summarizes visual inputs; native-video API models stream the clip directly, while local vLLM models sample frames at the listed `video_fps` or fixed count. For image-form MAR questions, option reference frames are appended as additional visual inputs. Token budgets are tight because MAR outputs a single letter; thinking variants retain a larger budget for the hidden reasoning trace that is stripped before choice extraction.

Models	Frames / video input	Temp.	Max new tokens
Closed-source / API models			
gemini-3-flash-preview; gemini-3.1-pro-preview	native video (Files API)	0.0	–
qwen3.5-omni-plus	native video	0.0	8192
kimi-k2.5	frames mode (default); `video` mode available	0.0	8192
kimi-k2.5 (thinking)	frames mode (default); `video` mode available	0.0	8192
gpt-4o	32 frames (JPEG, uniform)	0.0	512
Open-source / local models			
Qwen3-Omni-30B-A3B-Instruct	native video; `video_fps`=0.25; audio muted	0.0	8192
Qwen3-Omni-30B-A3B-Thinking	native video; `video_fps`=0.25; audio muted	0.6	8192
MiniCPM-V-4_5	32 frames cap; `video_fps`=0.25	0.0	512
Qwen2.5-VL-7B-Instruct	native video; `video_fps`=0.5	0.0	5
Qwen3-VL-8B-Instruct	native video; `video_fps`=0.5	0.0	5
Qwen3-VL-8B-Thinking	native video; `video_fps`=0.5	0.0	8192
LongVA-7B-DPO	96 frames	0.0	512
InternVL3-8B	96 frames (pre-resized 448×448)	0.0	5
InternVL3_5-8B-Instruct	150 frames (pre-resized 448×448)	0.0	5
InternVL3_5-8B-Instruct (thinking)	150 frames (pre-resized 448×448)	0.6	5
llava-onevision-qwen2-7b-ov-hf	48 frames	0.0	512
LongVU_Qwen2_7B	48 frames	0.0	512
InternVL2_5-8B	40 frames (tiled via `max_dynamic_patch`)	0.0	5
LLaVA-Video-7B-Qwen2	48 frames	0.0	512

Table 38: MLI label distribution.

Label	#Samples	Percent
within_24h	46	25.8%
more_than_24h	132	74.2%

Table 39: Detailed statistics for MLI video clips and annotations.

Metric	**Mean (μ)**	**Median**	**Min**	**Max**
Annotations per Participant	11.87	12.0	0	44
Video Clip Duration (s)	14.79	15.0	5.6	31.4

Table 40: MLI inference settings. "Frames/input" summarizes visual inputs; native-video vLLM models sample internally (e.g. Qwen2.5/3-VL). Gemini uses full video via the Files API without a fixed generation-token cap in the wrapper. Kimi defaults to `mode=video` (full video); `n_frames` applies only in `frames` mode.

Models	Frames / video input	Temp.	Max new tokens
Closed-source / API models			
gemini-3-flash-preview; gemini-3.1-pro-preview	native video	0.0	–
qwen3.5-omni-plus	native video	0.0	4096 ($\max(\cdot, 10)$ in code)
kimi-k2.5	native video	0.0	8192
kimi-k2.5 (thinking)	native video	0.0	8192
gpt-4o	32 frames (JPEG)	0.0	4096
Open-source / local models			
Qwen3-Omni-30B-A3B-Instruct	native video	0.0	4096
Qwen3-Omni-30B-A3B-Thinking	native video	0.6	8192
MiniCPM-V-4_5	native video; up to 180 frames / slot, 3 slots; `video_fps`=1	0.0	4096
Qwen2.5-VL-7B-Instruct	native video; `video_fps`=1.0	0.0	4096
Qwen3-VL-8B-Instruct	native video; `video_fps` default	0.0	4096
Qwen3-VL-8B-Thinking	native video; `video_fps` default	0.0	8192
LongVA-7B-DPO	360 frames	0.0	4096
InternVL3-8B	150 frames	0.0	4096
InternVL3_5-8B-Instruct	96 frames	0.0	4096
InternVL3_5-8B-Instruct (thinking)	96 frames	0.6	8192
llava-onevision-qwen2-7b-ov-hf	64 frames	0.0	4096
LongVU_Qwen2_7B	64 frames; early stop on final label line	0.0	2048
InternVL2_5-8B	58 frames	0.0	4096
LLaVA-Video-7B-Qwen2	64 frames	0.0	4096

distinct capabilities: verbal/event-level inference for text questions and fine-grained visual grounding for image questions. Among closed-source/API models, Gemini-3-flash-preview obtains the best text-form macro F1 (51.39) and the best text-form accuracy (53.85), while Gemini-3.1-pro-preview leads on the image form with the best macro F1 (63.21) and accuracy (64.10); the two Gemini variants thus trade off between the two MAR forms, suggesting that text-form MAR benefits from faster decoding and image-form MAR benefits from stronger visual grounding rather than from a single dominant model. Among open-source/local models, InternVL3_5-8B-Thinking reaches the best text-form macro F1 (52.83) and accuracy (56.86)—in fact surpassing every closed-source entry on the text form—whereas InternVL2_5-8B attains the best image-form macro F1 (41.38) and accuracy (46.15). The per-form scores also reveal a systematic weakness on image-form MAR for several local video models: MiniCPM-V-4_5 (10.26), LLaVA-Video-7B-Qwen2 (9.57), LongVA-7B-DPO (15.48), and llava-onevision-qwen2-7b-ov-hf (14.38) all fall well below the 25% random-choice level in macro F1, indicating that appending bbox-anchored option frames as extra visual inputs is not natively supported by these architectures.

Despite Gemini-3.1-pro-preview reaching 63.21 Macro F1 on image-form MAR and Gemini-3-flash-preview reaching 51.39 on text-form, gaps of roughly 3.5 and 9.7 points to human performance (66.67 / 61.11) remain, widening further in accuracy (10.9 points on image-form, 6.2 points on text-form). The difficulty concentrates in cases where the item worth remembering is identified not by visual prominence or verbal explicitness alone but by the user's momentary intent—encoded in subtle gaze dwell, object-proximity cues, and task context—signals that current VLMs cannot reliably infer from a short egocentric clip with a fixation overlay. For all results, see Table 31.

MLI is evaluated primarily with macro F1 because the binary labels are moderately imbalanced and a useful memory assistant must recognize both short-term information and long-term ones. Among closed-source/API models, Gemini-3.1-pro-preview obtains the best macro F1 (82.34) and the best acc(85.39). Among open-source/local models, InternVL3-8B reaches the best macro F1 (71.86)and the best acc(75.84). The per-class scores reveal that LongVU, LLaVA-onevision, and LongVA strongly favor the `within_24h` label. For all results, see Table 31

The evaluation of the human expert baseline on MLI reveals a profound gap between human cognitive precision and current MLLM capabilities. As shown in Table 31, human experts achieve an exceptional **accuracy of 88.89%** and a **macro F1-score of 87.50%**, demonstrating a robust ability to perform long-term temporal grounding. A more granular analysis of the F1-scores indicates that human experts maintain high reliability across different time scales, scoring **83.33% for short-term recall** ($\leq 24h$) and an impressive **91.67% for long-term recall** ($> 24h$). In stark contrast, even the top-performing closed-source models, such as *Gemini-3.1-pro-preview*, exhibit a significant performance drop, particularly in short-term precision.

L.3.2 Ablation Study

Table 41 presents the ablation study results for the *Memory Assistance Recognition* (MAR) task, evaluating the performance of both closed-source (*Kimi-k2.5*) and open-source (*Qwen3-VL-8B-Instruct*) models under various input configurations. Because MAR contains two question forms with distinct input structures—*text-form* questions (four textual candidates, primarily from short-term cues) and *image-form* questions (four bbox-anchored reference-frame candidates, for long-term object memory)—we report each form separately. The analysis provides several critical insights into how multimodal cues and prompting strategies contribute to modeling memory-assistance intent:

- **The Pivotal Role of ICL on Image-form Questions:** The integration of *In-Context Learning* (ICL)—where a single demonstration question is injected as a textual few-shot example, with one Gemini-2.5-Pro-generated sentence per bbox-anchored option—emerges as a primary driver of performance gains on image-form questions. For *Kimi-k2.5*, adding ICL to the video and gaze baseline yields a significant improvement, increasing Accuracy from 29.03% to 38.71% and Macro F1 from 18.61 to 28.75. This suggests that a grounded textual description of each candidate object provides the semantic anchor the model needs to disambiguate visually similar distractors within the same scene.
- **Observations on Gaze Information:** The isolated addition of gaze signals g_i to the video-only baseline has a form-dependent effect. On text-form questions, gaze consistently hurts both models: *Kimi-k2.5*'s accuracy drops from 52.50% to 47.50% (Macro F1 48.08 $\rightarrow$ 42.02) and *Qwen3-VL-8B-Instruct*'s from 40.00% to 35.00%. On image-form questions, in contrast, gaze is a genuine asset, raising *Qwen3-VL-8B-Instruct*'s accuracy from 29.03% to 32.26% (Macro F1 25.83 $\rightarrow$ 28.61). This asymmetry suggests that raw gaze duration is informative when the task reduces to spatial localisation of an intended object, but acts as a visual distractor when the decision has to be made by linguistic reasoning over short-term cues.
- **Mixed Behaviour of ICL on Text-form Questions:** On text-form questions, the benefit of ICL is no longer uniform. *Qwen3-VL-8B-Instruct* recovers to its video-only baseline under the full configuration (40.00% Accuracy, 37.67 Macro F1), while *Kimi-k2.5*'s accuracy keeps rising to 55.00% but its Macro F1 collapses from 42.02 to 24.93, indicating that the model's predictions concentrate on a single dominant option letter rather than calibrating to the participant's memory pattern. This asymmetry suggests that when short-term memory cues are linguistically heterogeneous across a participant's items, a single demonstration can encourage shortcut mimicry instead of genuine generalisation, particularly in higher-capacity API models.
- **Model Comparison:** Across all configurations, the closed-source *Kimi-k2.5* maintains a clear lead on text-form questions, peaking at 55.00% Accuracy against the 40.00% Accuracy of *Qwen3-VL-8B-Instruct*. On image-form questions the two models are much closer once ICL is introduced (Macro F1 28.75 vs. 29.09), indicating that the fine-grained visual grounding required for image-form MAR is not a capability that scales trivially with model size and benefits comparably from an additional textual grounding cue.

In summary, the results underscore that ICL is indispensable for bridging the gap between raw egocentric video v_i and object-level memory-assistance intent on image-form questions, while gaze signals offer selective value only when the underlying task is spatial rather than linguistic.

Table 42 presents the performance of models in predicting the temporal utility of memory items, specifically distinguishing between short-term ($\mathcal{S}$, within 24 hours) and long-term ($\mathcal{L}$, beyond 24

Table 41: Ablation results for MAR. All three input settings are evaluated on the same 71-question ICL-comparable subset (40 text-form + 31 image-form); every model produces a valid answer on every question. Bold indicates the best performance within each model category.

Model	Inputs	Overall		Text-form		Image-form	
		Acc	Macro F1	Acc	Macro F1	Acc	Macro F1
Closed-source/API models							
Kimi-k2.5	Video	40.85	**37.21**	52.50	48.08	25.81	19.68
	Video + Gaze	39.44	33.44	47.50	42.02	29.03	18.61
	Video + Gaze + ICL	**47.89**	30.40	**55.00**	24.93	**38.71**	**28.75**
Open-source/local models							
Qwen3-VL-8B-Instruct	Video	35.21	33.30	**40.00**	**38.36**	29.03	25.83
	Video + Gaze	33.80	30.94	35.00	32.60	32.26	28.61
	Video + Gaze + ICL	**36.62**	**33.83**	**40.00**	37.67	**32.26**	**29.09**

Table 42: Ablation results for MLI. Bold indicates the best performance within each model category.

Model	Inputs	MLI			
		Acc	Macro F1	F1($\leq 24h$)	F1($> 24h$)
Closed-source/API models					
Kimi-k2.5	Video	76.97	71.99	60.19	83.79
	Video + Gaze	**78.09**	**74.89**	**65.49**	**84.30**
Open-source/local models					
Qwen3-VL-8B-Instruct	Video	**59.55**	**58.67**	**52.63**	**64.71**
	Video + Gaze	46.07	45.96	48.39	43.53

hours) retention needs. The results reveal distinct behaviors in how closed-source and open-source models process physiological and visual cues for lifespan prediction:

- **Effective Multi-modal Integration in Kimi-k2.5:** For the *Kimi-k2.5* model, the inclusion of gaze signals g_i alongside video v_i leads to a consistent performance improvement across all metrics. The Accuracy increases from 76.97% to 78.09%, with a notable gain in the Macro F1 score (from 71.99 to 74.89). This enhancement is particularly pronounced in the $\mathcal{S}$ class (F1 score rising from 60.19 to 65.49), suggesting that fine-grained physiological signals such as gaze assist the model in identifying the immediate, transient nature of certain user intents.

- **Performance Degradation in Open-source Models:** In contrast, *Qwen3-VL-8B-Instruct* exhibits a significant performance drop when gaze information is introduced, with Accuracy falling from 59.55% to 46.07%. This discrepancy indicates that while open-source models possess basic visual recognition capabilities, they struggle to effectively align and interpret raw physiological signals for high-level cognitive tasks like lifespan identification, potentially due to the lack of specialized training on synchronized egocentric-physiological datasets.

- **Class-specific Bias and Information Importance:** Across both models, performance on long-term memory items ($\mathcal{L}$) is generally superior to that on short-term items ($\mathcal{S}$). This suggests that information requiring long-term retention—such as *semantic knowledge* or *meaningful experiences* that users intend to preserve—may possess more distinct contextual signatures than the ephemeral, state-based information in the $\mathcal{S}$ category. However, transient information often carries higher *immediate utility* and mission-critical importance in daily life (e.g., the temporary location of an object). The observed performance gap underscores the need for more targeted training and refined feature extraction to better capture the subtle cues associated with short-term memory needs.

In conclusion, the results demonstrate that while predicting memory lifespan is a challenging cognitive task, the integration of gaze signals can provide valuable cues for identifying information utility, provided the model has sufficient capacity to fuse these heterogeneous modalities.

M ANNOTATION RELIABILITY ANALYSIS AND TEST-RETEST RESULTS

Given that annotations in EgoIntrospect describe subjective internal human states provided directly by participants, assessing their temporal stability and reliability is crucial for establishing ground-truth validity. To evaluate label stability across identical experiential episodes, we conducted a rigorous *test–retest study* across a multi-month interval.

M.1 DATA COLLECTION PROTOCOL

- **Participants**: A random subset of 6 original participants was invited, among whom 5 participants (P1, P2, P7, P15, P44) completed full re-annotation passes within the requested timeframe.
- **Temporal Interval**: The re-annotation was conducted more than 6 months after the initial recording and annotation sessions.
- **Blinded Protocol**: Participants were provided with the original egocentric video clips while all previous annotation labels were strictly withheld. They reviewed each video episode and re-entered their internal states and annotations.
- **Evaluation Scale**: In total, the re-test study yielded 644 field-level comparative pairs across tasks.

M.2 AGREEMENT METRICS

We employed three complementary comparators depending on the target data type:

1. **Semantic Similarity (Free Text)**: Evaluated via an LLM judge (`Kimi-K2.5`, temperature 0). Two free-text descriptions are deemed in agreement if they refer to the same event, intention, or underlying meaning. Paraphrases, language variations, and differing levels of detail count as agreement, whereas non-overlapping events or intents count as disagreement.
2. **Spatial Overlap (Bounding Boxes)**: Evaluated for spatial grounding fields. Two bounding boxes agree if their Intersection-over-Union (IoU) ≥ 0.15 or if their spatial centers lie within half the mean bounding box dimension.
3. **Exact Match (Categorical/Ordinal)**: Applied to closed-vocabulary fields and ordinal rating scales (e.g., discrete intensity ranges).

For multi-item sets (e.g., Task 2.2 Accepted Recommendations), text items were paired greedily using semantic similarity and scored as $\frac{\sum \text{similarity}}{\max(|A|,|B|)}$. For document-level recall (Task 3.1), semantic similarity was computed holistically over the concatenated text without item-level pairing.

M.3 TEST-RETEST RESULTS

Table 43 summarizes the agreement scores across two complementary aggregation perspectives:

- **Per-Participant Mean $\pm$ SD**: Averages the agreement rates across the five participants individually, ensuring equal weighting per subject and capturing cross-subject variation.
- **Pooled Average**: Computes overall agreement across all individual scored instances (n).

M.4 DISCUSSION AND QUALITATIVE INTERPRETATIONS

The empirical results demonstrate robust reliability for in-situ anchored tasks (Tasks 1.1 and 2.1), where agreement exceeds 90% across spatial, semantic, and textual fields after more than six months. This confirms that participants reliably recall their primary salient targets, intentions, and underlying rationales when presented with the original video context.

For more subjective or dynamic constructs, we observe greater variance across the multi-month gap:

- **Trait-like vs. State-like Constructs**: Psychometric test–retest stability assumes stable, trait-like targets. However, internal user states like emotional intensity (`emotion_level`, 55.6%), expected assistant responses (68.9%), and accepted recommendations (56.5%) are inherently state-like, sensitive to real-time context, and subject to natural memory decay over a 6-month period. Because participants evaluate past events without the exact real-time affective states and cognitive

Table 43: **Test–retest agreement scores across annotation fields over a 6-month interval.** Two complementary averages are reported: per-participant mean ± standard deviation and overall pooled agreement (n denotes total comparison count).

Task	Field / Target	Comparator	Per-Participant Mean ± SD (%)	Pooled (%)	*n*
Task 1.1 (MoR)	Bounding Box (Object to Capture)	Same Object	95.6 ± 6.5	96.4	55
	`explanation` (Recording Rationale)	Semantic	98.2 ± 3.6	98.6	69
Task 1.2 (Emotion)	`description` (Episode Description)	Semantic	92.5 ± 11.2	96.1	51
	`emotion_label`	Semantic	81.1 ± 19.6	85.2	54
	`emotion_level` (Scale 1–5)	**Exact**	$\mathbf{55.8 \pm 16.0}$	**55.6**	54
	Bounding Box	Same Object	91.7 ± 14.4	94.1	17
Task 2.1 (Visual Q&A)	`text` (User Question)	Semantic	93.6 ± 10.8	95.0	101
	`expected_response`	Semantic	$\mathbf{61.9 \pm 34.8}$	**68.9**	61
	Bounding Box	Same Object	100.0 ± 0.0	100.0	22
Task 2.2 (Rec. Acceptance)	**Accepted Recommendations (Set)**	Semantic Set	$\mathbf{54.8 \pm 25.2}$	**56.5**	112
Task 3.1 (Episodic Recall)	Events & Vivid Details (Doc-Level)	Semantic	100.0 ± 0.0	100.0	5
Task 3.2 (Memory Intent)	Short- vs. Long-Term Horizon	Exact	86.7 ± 18.3	86.7	15
	`event_context`	Semantic	83.3 ± 33.3	83.3	12
	`memory_reason`	Semantic	80.0 ± 29.8	80.0	15

loads present during recording, minor variations across these dynamic fields are fully expected and reflect natural hindsight re-framing rather than annotation noise.

- **Contextual Evolution vs. Noise**: Follow-up debriefing interviews were conducted by revealing the original baseline annotations to participants. Upon review, participants demonstrated high subjective agreement with their initial labels across emotional level (91.7%), expected response (95.1%), and recommendation acceptance (97.5%).

Participants noted that re-test variances stemmed primarily from evolving situational context rather than annotation noise: an information gap present six months prior may have since been resolved, or immediate task priorities may have shifted. Consequently, variance in these state-like fields reflects the natural dynamism of human internal experiences, confirming the overall soundness and repeatability of our annotation protocol.

N Cross-Task Interdependence Study

To evaluate the value of capturing multiple internal-state dimensions within shared user sessions, we investigate the underlying cross-task synergies from two perspectives: *cross-task temporal overlap* and *cross-task transfer dynamics*.

N.1 Motivation and Conceptual Framework

EgoIntrospect considers affect, interaction intent, and memory as complementary dimensions of a user-specific internal state that evolves continuously over time. A wearable AI assistant must reason jointly about these internal states to provide personalized and proactive assistance. Collecting these annotations on a shared per-user timeline enables within-user analyses of how different internal states co-occur, diverge, and jointly characterize user experiences. This design follows the established paradigm of unified egocentric datasets (e.g., Ego4D) that leverage shared user recordings to support multi-task evaluation.

N.2 Cross-Task Temporal Analysis

To determine whether annotations across different tasks naturally cluster around shared salient periods, we performed a temporal overlap analysis on the continuous recording timeline of the test set.

Specifically, we aligned four representative annotation tasks along each participant's continuous timeline: Moment of Recording (MoR, Task 1.1), Affective State (Task 1.2), Vivid Memory Details (Task 3.1), and Memory Assistance Need (Task 3.2). For each task, we constructed a Gaussian kernel-smoothed temporal intensity distribution, normalized each curve to unit area, and measured its pairwise overlap with other tasks. Statistical significance was evaluated using two complementary null models:

1. **Within-session circular-shift null**: Preserves the intrinsic temporal clustering of each task while disrupting pairwise alignment via a random circular shift along the session timeline.
2. **Cross-user mismatch null**: Pairs annotations across different participants, thereby controlling for general dataset-wide temporal biases.

As summarized in Table 44, all task pairs exhibit statistically significant temporal overlap ($p < 0.01$ or $p < 0.001$, adjusted using Benjamini-Hochberg FDR) compared to both null baselines. These results confirm that annotations across tasks are not independently distributed; rather, moments users wish to record, emotionally salient experiences, vivid recall points, and information requiring memory assistance consistently cluster around shared, personally salient periods.

Table 44: **Pairwise temporal overlap among tasks and baseline null significance.** Significance levels are based on p-values adjusted for BH-FDR ($^{**}p < 0.001$, $^{*}p < 0.01$). All task pairs show significant temporal alignment compared to both circular-shift and cross-user null distributions.

Task Pair	Observed Overlap	Circular-Shift Null	Shift p	Cross-User Null	Cross-User p
1.1 – 1.2	0.330	0.232	$< .001^{**}$	0.243	$.001^{**}$
1.1 – 3.1	0.360	0.250	$< .001^{**}$	0.262	$.001^{**}$
1.2 – 3.1	0.371	0.233	$< .001^{**}$	0.245	$.001^{**}$
1.1 – 3.2	0.281	0.184	$< .001^{**}$	0.180	$.001^{**}$
3.1 – 3.2	0.281	0.181	$< .001^{**}$	0.186	$.001^{**}$
1.2 – 3.2	0.237	0.172	$.007^{*}$	0.171	$.002^{*}$

N.3 Cross-Task Transfer Analysis

To directly evaluate whether user-specific context from one task paradigm can benefit reasoning in another, we conducted cross-task in-context transfer experiments. Specifically, for a given target Task_B, we selected an annotation sample from the same user's Task_A as an in-context exemplar ($\mathrm{Task}_A \rightarrow \mathrm{Task}_B$) to test if cross-task context improves performance over the single-task baseline.

The cross-task transfer performance and relative accuracy gains across key target tasks are summarized in Table 45.

Table 45: **Cross-task in-context transfer performance and relative accuracy gains.** Injecting user-specific exemplars from complementary source tasks ($\mathrm{Task}_A \rightarrow \mathrm{Task}_B$) consistently outperforms single-task baselines.

Target Task	Single-Task Baseline Acc.	Top Transfer Sources ($\mathrm{Task}_A \rightarrow \mathrm{Task}_B$)	Transfer Acc.	Relative Gain (Δ)
Task 1.1 (Recording)	**53.97%**	**Task 2.1 → Task 1.1**	**74.60%**	**+20.63%**
		Task 2.2 → Task 1.1	66.67%	+12.70%
		Task 3.1 → Task 1.1	65.08%	+11.11%
		Task 3.2 → Task 1.1	63.49%	+9.52%
Task 2.1 (Intent)	**46.67%**	**Task 1.1 → Task 2.1**	**53.33%**	**+6.66%**
Task 3.1 (Memory Recall)	**42.86%**	**Task 2.1 → Task 3.1**	**57.14%**	**+14.28%**
		Task 2.2 → Task 3.1	**57.14%**	**+14.28%**
		Task 1.1 / 1.2 → Task 3.1	50.00%	+7.14%

Key Insights from Transfer Results

1. **Substantial Cross-Task Gains for Recording & Memory**: Incorporating interaction intent context (Task 2.1) into Task 1.1 (Moment of Recording) yielded the most significant performance jump, boosting accuracy from 53.97% to 74.60% (+20.63% net gain). Similarly, providing intent context (Tasks 2.1, 2.2) consistently enhanced cognitive memory recall (Task 3.1), elevating accuracy from 42.86% to 57.14% (+14.28% gain). These findings empirically demonstrate that a user's underlying intent and memory context serve as strong predictive priors for their recording desires, validating the necessity of capturing all tasks within shared user sessions.
2. **Broad Transfer Synergy Across Tasks (21/30 Positive or Neutral)**: Among all 30 evaluated cross-task pairs, 10 pairs achieved noticeable accuracy improvements while 11 pairs maintained stable baseline performance, indicating that cross-task context generally acts as a constructive prior without causing negative interference.

Together, the temporal overlap statistics and transfer experiments demonstrate that collecting these multi-dimensional internal-state annotations from the same users within shared sessions provides unique contextual synergies that isolated single-task datasets cannot offer.

O LIMITATION

Despite the heavy investments, EgoIntrospect has several limitations. Although our dataset captures richer human-centered signals than prior egocentric benchmarks, the scale of participant diversity remains limited compared to internet-scale datasets. Compared to prior pioneer works such as Ego4D, the demographic diversity of our dataset remains relatively limited due to the substantial cost and logistical challenges associated with large-scale cross-national multisensory data collection and annotation.

On the modeling side, while we provide comprehensive analyses of existing models and explore simple strategies such as in-context calibration of individual differences, we do not propose a dedicated modeling framework specifically designed for human-centered internal-state understanding. Yet, we hope our work can motivate future research on architectures that better model subjective user states, long-term personalization, and individual behavioral patterns.